Modelos de Regressão Linear para Variáveis Intervalares

Uma extensão do modelo ID

por

Pedro Jorge Correia Malaquias

Dissertação do Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Faculdade de Economia do Porto

Orientada por: Professora Dra. Sónia Dias
Coorientada por: Professora Dra. Maria Paula Brito

Setembro, 2017

Dedicado a ti, avô.

Nota Bibliográfica

Pedro Malaquias nasceu na cidade de Aveiro em 1994.

Concluiu a licenciatura em Gestão em 2015 na Faculdade de Economia da Universidade do Porto e neste mesmo ano ingressou o Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão na mesma instituição.

Profissionalmente exerce funções numa empresa de retalho, onde tem como principais atividades analisar a informação do negócio e trabalhar em projetos diretos na sua área de atuação de forma a trazer valor para a empresa.

Agradecimentos

Tendo chegado ao final deste trabalho, cabe-me agradecer a todas as pessoas que foram importantes na sua elaboração e as que me apoiaram nestes últimos anos:

À minha namorada Inês pela paciência, compreensão e motivação.

À minha mãe e ao meu pai que possibilitaram os meus estudos noutra cidade e a todo o apoio e motivação demonstrados.

À minha avó, às minhas tias e a toda a família por toda a ajuda sempre que foi necessário.

Às chefias do meu departamento pelo o apoio e a flexibilidade para a conclusão deste desafio e aos restantes colegas da DPA, nomeadamente pela cedência dos dados.

Um especial agradecimento à professora Sónia Dias e à professora Paula Brito pela paciência, sugestões e melhorias críticas ao longo de todo o trabalho, tornando este resultado final possível.

À professora Paula Cheira e ao professor Pedro Silva pelo apoio na implementação do modelo em *R software*, contribuindo com algumas das funções paralelas utilizadas.

Resumo

A Análise de Dados Simbólicos permite captar a variabilidade existente nos dados quando estes são agregados a níveis superiores aos seus registos, respondendo a variadas necessidades estatísticas e tornando-se numa ferramenta mais poderosa do que a Análise de Dados Clássica. No contexto das variáveis simbólicas intervalares foram propostos vários métodos de regressão linear. Dias e Brito (2017) propuseram um método que utiliza funções quantil, com uma distribuição inerente, para a representação destes intervalos e posterior aplicação de uma regressão linear. Neste trabalho este modelo foi adaptado de forma a considerar uma distribuição Triangular Geral, o que permitiu demonstrar a versatilidade do Modelo e incrementar o seu desempenho. O modelo desenvolvido por Dias e Brito (2017) e este novo desenvolvimento foi implementado em *R software*.

Abstract

Symbolic data analysis allows us to capture the variability of data when they are aggregated at levels higher than their observations. These methods respond to varying statistical needs and become a more powerful tool that the Classical Data Analysis methods. In the context of interval symbolic variables, several linear regression methods were proposed. Dias and Brito (2017) proposed a method that uses quantile function, with an inherent distribution, for the representation of these intervals and later application of a linear regression. In this work, this model was adapted to consider a Triangular distribution, which allowed to demonstrate the versatility of the model and to increase its performance. The model developed by Dias and Brito (2017) and its adaptation to the context of interval variables assuming a Triangular distribution was implemented in *R software.*

Conteúdo

Capítulo 1: Introdução 1

1.1 Motivação 1

1.2 Problema a estudar 2

1.3 Organização da Dissertação 2

Capítulo 2: Dados Simbólicos: uma breve introdução 4

2.1 Obter dados simbólicos a partir de dados clássicos 5

2.2 Variáveis Simbólicas 6

Capítulo 3: Estado da Arte 9

3.1 Modelos de regressão linear para variáveis intervalares 10

3.1.1 Método do Centro 10

3.1.2 Método do Mínimo e Máximo 12

3.1.3 Método do Centro e Raio 13

3.1.4 Método do Centro e Raio com Restrições 15

3.1.5 *Lasso IR-Method* 16

3.1.6 *Interval Distribution (ID) Regression Model* 18

Capítulo 4: ID Regression Model: Novas Abordagens 22

4.1 *ID Model* e a representação dos intervalos por funções quantil 22

4.2 Distância de *Mallows* 28

4.3 Medidas de Avaliação do Modelo 32

4.4 *ID Regression Model*: Distribuição Uniforme 33

4.5 *ID Regression Model*: Distribuição Triangular Simétrica 36

4.6 *ID Regression Model*: Distribuição Triangular Geral 38

Capítulo 5: Implementação e Aplicação do Modelo 44

5.1 Implementação do Modelo 44

5.2 Aplicação do modelo a um caso real 48

Capítulo 6: Conclusões 58

Bibliografia 60

Anexos: Código R 63

Lista de Figuras

4.1 Representação de um intervalo, do seu intervalo simétrico e respetivas funções quantil, assumindo uma distribuição Uniforme em cada intervalo……24

4.2 Representação de um intervalo, do seu intervalo simétrico e respetivas funções quantil, assumindo uma distribuição Triangular Simétrica em cada intervalo..……...…...26

4.3 Representação de um intervalo, do seu intervalo simétrico e respetivas funções quantil, assumindo uma distribuição Triangular Geral em cada intervalo… 28

5.1 Exemplo de uma tabela de dados clássicos, para transformação em intervalos pela função "*ToSimbolic*" em *R software*……... 45

5.2 Exemplo de uma matriz "input" de dados simbólicos……... 46

5.3 Variável Resposta Consumo (*kWh*) no formato de uma tabela simbólica… 49

5.4 Representação dos intervalos observados e previstos no mês de agosto…... 50

5.5 Representação dos intervalos observados e previstos no mês de novembro. 51

5.6 Representação dos intervalos observados e previstos para um período quente…….54

5.7 Representação dos intervalos observados e previstos para um período frio….…….55

Lista de Tabelas

2.1 Informação do serviço de atendimento bancário …….. 4

2.2 Dados Simbólicos: Informação do serviço de atendimento bancário ……5

5.1 Medidas de avaliação do modelo ……..50

5.2 Relações lineares para cada um dos modelos: aplicação anual……..... 52

5.3 Medidas de Qualidade para cada um dos modelos: aplicação anual……. 52

5.4 Medidas de avaliação do modelo ……..54

5.5 Relações Lineares para cada um dos modelos: aplicação período Quente e período Frio …….…….. 55

5.6 Medidas de Qualidade para cada um dos modelos: aplicação período Quente e período Frio …….. 56

Capítulo 1

Introdução

Com a evolução tecnológica dos últimos anos, existe uma necessidade crescente de desenvolver novos métodos de análise que sejam capazes de explorar e compreender dados complexos. Esta crescente complexidade da informação disponível caracteriza-se por volumosas bases de dados que requerem métodos de recolha dos dados mais eficazes.

Os métodos clássicos de análise consideram dados representados em tabelas clássicas, onde cada célula contém o valor do atributo para uma observação. No entanto, estes métodos não correspondem às necessidades atuais para a resolução de problemas, sendo limitados quando as análises a efetuar incidem sobre dados com variabilidade (Brito 2014).

Têm sido desenvolvidos métodos para representar e analisar os dados, de forma a ser possível abordar problemas mais complexos onde é necessário realizar análises a níveis superiores ao dos dados recolhidos. Uma área que se mantem em constante desenvolvimento é a Analise Simbólica de Dados, que permite ultrapassar a limitação dos métodos clássicos na análise de dados com variabilidade, através da agregação dos valores individuais em variáveis simbólicas que caracterizam cada "grupo" (Brito e Noirhomme-Fraiture 2006).

1.1 Motivação

Com os desenvolvimentos da análise simbólica de dados torna-se possível realizar análises a dados mais complexos e volumosos, em que a unidade estatística de interesse é de nível superior às observações, considerando a variabilidade inerente dos mesmos. A investigação de novos métodos teóricos irá ser fundamental para desenvolvimentos nesta área, com o objetivo de suportar a crescente complexidade de estudos científicos e económicos.

A análise simbólica revolucionou a forma como diferentes tipos de dados agregados são analisados. A agregação e análise de variáveis contínuas poderá ser realizada através de intervalos, assumindo uma distribuição inerente aos mesmos. Nos modelos e métodos atualmente desenvolvidos é assumido que a distribuição inerente às variáveis intervalares é a distribuição Uniforme, sendo que a consideração de novas distribuições poderá marcar um desenvolvimento importante nesta área.

1.2 Problema a estudar

Nesta dissertação irá ser estudado e implementado um desenvolvimento que irá contribuir para demostrar a versatilidade do Modelo de Regressão Linear designado por *Interval Distributional (ID) Regression Model* proposto por Dias e Brito (2017).

O *ID Regression Model* considera os valores das variáveis simbólicas intervalares representados por uma função quantil, função inversa da função de distribuição acumulada, assumindo uma distribuição em cada intervalo observado. Em geral, é assumida a distribuição Uniforme e esta hipótese serviu como base para os desenvolvimentos estatísticos realizados até ao momento. A versatilidade deste modelo, no entanto, possibilita considerar outras distribuições além da distribuição Uniforme, e nesta dissertação irá ser considerada a Distribuição Triangular. O novo modelo obtido será testado e comparado com outros modelos propostos na literatura. Será desenvolvido um pacote no *R software* para a sua implementação.

1.3 Organização da Dissertação

Esta dissertação é estruturada em seis capítulos. No primeiro capítulo são abordadas várias considerações sobre o tema a desenvolver, como a motivação e o problema a estudar. No segundo capítulo serão apresentados, no âmbito da Análise Simbólica de Dados, os principais conceitos e definições e os vários tipos de variáveis simbólicas. No terceiro capítulo será realizada uma revisão bibliográfica de métodos de regressão linear para variáveis intervalares, incluindo o modelo proposto por Dias e Brito (2017). No quarto capítulo é apresentada uma proposta para a extensão do *ID Model* para variáveis

intervalares, onde a distribuição inerente a esses intervalos é a distribuição Triangular (Dias e Brito 2017). No quinto capítulo o novo modelo será testado e avaliado, sendo realizado uma comparação com os restantes modelos propostos na literatura, utilizando dados reais. Para realizar os testes e comparações irá ser utilizado o *R software* e pacotes associados. No sexto capítulo serão apresentadas as principais conclusões do trabalho desenvolvido, os principais problemas identificados e o trabalho a ser desenvolvido no futuro.

Capítulo 2

Dados Simbólicos: uma breve introdução

Em Estatística e análise multivariada de dados clássicos, as análises realizadas incidem sobre indivíduos singulares descritos por um conjunto de variáveis numéricas ou categóricas.

Na Tabela 2.1 é apresentado um exemplo de dados no formato clássico sobre o serviço de atendimento bancário em vários bancos. Em cada linha da tabela é apresentado para cada cliente, as suas características, o seu banco e o tempo médio de espera para ser atendido (em minutos). Quando é necessário analisar estes dados de forma a comparar o desempenho do atendimento e o perfil de clientes pela metodologia clássica, a informação é resumida a um único valor para cada instituição bancária, como uma média ou moda. No entanto, ao reduzirmos os dados de cada instituição bancária a um valor, a variabilidade existente é perdida.

Tabela 2.1: Informação do serviço de atendimento bancário

Cliente	Banco	Género	Idade	Educação	Rendimento Mensal	Temo de Espera (minutos)
Cliente 1	A	M	44	9º ano	1540	11
Cliente 2	B	F	70	12º ano	1000	37
Cliente 3	B	F	42	12º ano	1200	22
Cliente 4	C	M	56	Ed. Sup.	3000	3
...	...	...	...	...	...	...

Na Tabela 2.2 os dados foram agregados pelo nível da instituição utilizando variáveis simbólicas, ou seja, construindo uma tabela simbólica de dados. As variáveis clássicas deram lugar a variáveis simbólicas através da agregação dos dados em intervalos, histogramas e variáveis categóricas modais. A abordagem da Análise simbólica de Dados permite considerar a variabilidade inerente aos dados no momento da análise.

Tabela 2.2: Dados Simbólicos: Informação do serviço de atendimento bancário

Banco	Género	Idade	Educação	Rendimento Mensal	Temo de Espera (minutos)
A	{F,1/2; M,1/2}	[24, 79]	{6ºano, 1/4; 9ºano, 1/4; 12ºano, 1/4; Ed.Sup., 1/4}	[600, 2200]	{[0,10[,0.1; [10,20],0.6; [20,30],0.3}
B	{F,1/3; M.2/3}	[19, 82]	{4ºano, 1/3; 9ºano, 1/3; 12ºano, 1/3}	[550, 1350]	{[10,20[,0.4; [20,30],0.5; [30,40],0.2]
C	{F,4/7; M.3/7}	[27, 93]	{12ºano, 1/3; Ed.Sup., 2/3}	[1000, 3500]	{[0,10[,0.7; [10,20],0.3]

A necessidade de analisar conjuntos de dados de maior dimensão e mais complexos levou ao desenvolvimento de métodos e modelos estatísticos que pudessem ser aplicados aos dados simbólicos.

A Análise Simbólica de Dados é uma abordagem que engloba um conjunto de métodos capazes de analisar grandes volumes informação que, em muitos casos, são analisados a um nível superior ao da sua recolha. A abordagem da análise simbólica de dados permite agregar os *microdados* em variáveis simbólicas, cujos "valores observados" são intervalos, histogramas ou distribuições de frequências, mantendo a variabilidade inerente aos registos. Por exemplo, os dados dos censos não devem ser analisados individualmente, por questões de privacidade, sendo que a Análise Simbólica de Dados permite analisar estes dados agregados por região, sem que a variabilidade em cada uma das regiões seja perdida (Brito e Noirhomme-Fraiture 2006).

2.1 Obter dados simbólicos a partir de dados clássicos

A forma mais comum para a transformação de dados clássicos em dados simbólicos é a agregação de observações individuais, sendo que é possível distinguir dois tipos de agregação (Brito 2014):

- **Agregação Temporal:** aplica-se quando os dados são recolhidos e observados ao

longo de um determinado período de tempo para o mesmo indivíduo ou entidade. Neste caso, os indivíduos são a unidade estatística de interesse para análise, sendo que a ordem temporal dos registos não é relevante, mas apenas os valores registados. Um exemplo de uma agregação temporal é o registo horário dos batimentos cardíacos de um paciente, sendo que estes devem ser agregados ao nível do dia, para cada um dos pacientes.

- **Agregação Contemporânea:** aplica-se quando os dados são recolhidos no mesmo momento do tempo e a unidade estatística de interesse encontra-se a um nível superior ao das observações. Os novos "grupos" são constituídos a partir de agregações dos valores individuais segundo características específicas. Um exemplo de uma agregação contemporânea é efetuada nos censos, onde os indivíduos inquiridos necessitam de ser agregados pela respetiva região.

Em determinadas situações, a Análise de Dados Simbólica pode trabalhar com bases de dados relativamente às quais não são conhecidos os *microdados*. Em alguns casos a informação a tratar, já associa a um mesmo indivíduo um conjunto ou intervalo de valores.

2.2 Variáveis Simbólicas

Para a representação de dados com variabilidade foram introduzidos novos tipos de variáveis que incluem mais informação e são mais complexas do que as variáveis clássicas, que registam apenas um valor ou categoria. À semelhança das variáveis clássicas, podemos também distinguir as variáveis simbólicas numéricas (ou quantitativas) e as categóricas (ou qualitativas). Na literatura, as variáveis clássicas numéricas e categóricas são consideradas um caso especial das variáveis simbólicas (Brito 2014).

Na definição das variáveis simbólicas apresentada em Brito e Noirhomme-Fraiture (2006), para um conjunto de variáveis $Y_1, \dots, Y_p$, O_j representa o domínio de Y_j e B_j o espaço de observação de Y_j, com $j = \{1, \dots, p\}$.

Variáveis quantitativas de valor único

Dado um conjunto de n observações $S = \{s_1, \dots, s_n\}$, uma variável quantitativa de valor

único Y é definida pela aplicação $Y: S \mapsto O$ tal que $s_i \mapsto Y(s_i) = \alpha \in O \subseteq IR$, com $i \in \{1, \dots, n\}$. Neste caso, B é idêntico ao conjunto subjacente *O*, ou seja, B ≡ *O*.

Variáveis categóricas de valor único

Dado $S = \{s_1, \dots, s_n\}$ e um conjunto finito de categorias $O = \{m_1, \dots, m_k\}$, uma variável categórica de valor único é definida pela aplicação $Y: S \mapsto O$, tal que $s_i \mapsto Y(s_i) = m_h$. Neste caso, $B \equiv O$. Se o conjunto O for ordenado, Y é uma variável ordinal, caso contrário, Y é uma variável nominal.

Variáveis quantitativas de valores múltiplos

Dado um conjunto S, uma variável quantitativa de valores múltiplos Y é definida por uma aplicação $Y: S \mapsto B$ tal que $s_i \mapsto Y(s_i) = \{\alpha_{i1}, \dots, \alpha_{in_i}\}$, onde B é o conjunto dos subconjuntos finitos de um conjunto subjacente $O \subseteq IR$. $Y(s_i)$ é assim um conjunto finito não vazio de números reais.

Variáveis intervalares

Sendo $S = \{s_1, \dots, s_n\}$, uma variável intervalar é definida pela aplicação $Y: S \mapsto B$ tal que $s_i \mapsto Y(s_i) = [l_i, u_i]$, onde B é o conjunto de intervalos de um conjunto subjacente $O \subseteq IR$. Um intervalo observado I_i pode ser definido pelos seus limites inferior, l_i, e superior, u_i, pelo ponto médio $c_{Y(s_i)} = \frac{u_i + l_i}{2}$ e pelo raio $r_{Y(s_i)} = \frac{u_i - l_i}{2}$, ou por uma função quantil $\Psi^{-1}_{Y(s_i)} = c_{Y(s_i)} + r_{Y(s_i)}(2t - 1)$, com $t \in [0,1]$.

Na literatura é, em geral, assumido que a distribuição inerente a cada intervalo é a distribuição Uniforme.

Exemplo: Considerando os dados simbólicos representados na Tabela 2.2, as variáveis "Idade" e "Salário Mensal" são um exemplo de variáveis intervalares.

Variáveis histograma

Quando a quantidade de valores (*microdados*) associados a cada uma das observações é elevada, a agregação em intervalos pode ser redutora. De forma a minimizar a perda de

informação pela agregação em intervalos é possível organizar essa informação em subintervalos e calcular as respetivas frequências.

Dado $S = \{s_1, \ldots, s_n\}$, uma variável histograma é definida por uma aplicação Y: S $\mapsto$ B, tal que $s_i \mapsto Y(s_i) = \{I_{i1}, p_{i1}; \ldots; I_{ik}, p_{ik}\}$ onde $I_{i1}, \ldots, I_{ik}$ são os subintervalos considerados para a observação s_i e $p_{i1}, \ldots, p_{ik}$ os pesos para cada um dos subintervalos, sendo que $p_{i1} + \ldots + p_{ik} = 1$. B é o conjunto de distribuições de frequências nos subintervalos, realçando-se que o número e amplitude dos subintervalos podem ser diferentes para cada observação. É assumido que para cada entidade s_i os valores são uniformemente distribuídos em cada subintervalo.

Exemplo: Considerando os dados apresentados na Tabela 2.2, a variável "Tempo de Espera (em minutos)" é um exemplo de uma variável histograma, em que cada intervalo relativo aos minutos em que cada cliente aguardou até ao seu atendimento tem um peso associado.

Variáveis categóricas multi-valor

Uma variável categórica multi-valor é definida pela aplicação Y: S $\mapsto$ B, tal que $s_i \mapsto Y(s_i) = \{m_{i1}, \ldots, m_{ip}\}$, onde B é o conjunto de subconjuntos finitos de $O = \{m_1, \ldots, m_k\}$.

Variáveis categóricas modais

Uma variável categórica modal Y com um domínio subjacente finito $O = \{m_1, \ldots, m_k\}$ é uma variável de valores múltiplos onde, para cada elemento, é registado um conjunto de l categorias, com $l \in \{1, \ldots, k\}$, a cada uma das quais está associada uma frequência ou probabilidade p_l. Neste caso, B é o conjunto de distribuições sobre O e os seus elementos são denotados por $Y(s_i) = \{m_{i1}, p_{i1}; \ldots; m_{ik}, p_{ik}\}$.

Exemplo: Considerando os dados apresentados na Tabela 2.2, a variável "Género" é um exemplo de uma variável categórica modal. A cada banco está associada a frequência dos clientes do sexo masculino e do sexo feminino.

Para simplificação da notação, nos capítulos seguintes vamos identificar a imagem, pela aplicação Y, da observação s_i, apenas por $Y(i)$. Nos casos das variáveis X_j, será usada a notação simplificada X_{ij}, em vez de $X_j(i)$.

Capítulo 3

Estado da Arte

Neste capítulo serão apresentados alguns dos principais modelos de regressão linear propostos para variáveis simbólicas intervalares.

O modelo de regressão linear é um modelo matemático usado para estudar a relação entre duas ou mais variáveis contínuas, no qual se tenta prever os valores de uma variável resposta, Y, através de um conjunto de variáveis explicativas, $X_1, \dots, X_p$. No modelo genérico de regressão linear, a representação matemática é dada por:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \tag{3.1}$$

sendo ε o termo residual.

De forma a estimar os parâmetros β_j, com $j \in \{1, \dots, p\}$ é utilizada uma estratégia de otimização, minimizando a soma dos quadrados dos resíduos (método dos mínimos quadrados). Estes parâmetros definem a relação entre a variável resposta Y e as variáveis explicativas X_j, com $j \in \{1, \dots, p\}$.

Nos últimos anos, foram propostos vários métodos de regressão linear para variáveis intervalares. Nas variáveis intervalares, os intervalos podem ser representados pelos seus centros e raios, extremos (superior e inferior) ou através de funções quantil. Esta última opção permite captar uma maior variabilidade dos dados, ao contrário das restantes formas de representação (Dias e Brito 2017).

As hipóteses probabilísticas que envolvem a teoria dos modelos de regressão linear para dados clássicos não serão consideradas neste contexto.

3.1 Modelos de regressão linear para variáveis intervalares

Considerando as variáveis intervalares, no contexto da Análise Simbólica de Dados, os modelos de regressão linear propostos, que se consideram mais relevantes são: o Método do Centro (Billard e Diday 2000); o Método do Mínimo e Máximo (Billard e Diday 2002); o Método do Centro e Raio (Lima Neto e De Carvalho 2008); o Método do Centro e Raio com Restrição (Lima Neto e De Carvalho 2010); o Lasso - *Constrained Regression Model* (Giordani 2014); e o *Interval Distribution (ID) Regression Model* (Dias e Brito 2017). Com estes métodos é possível estimar uma variável intervalar resposta a partir de p variáveis intervalares explicativas.

Os métodos atualmente propostos e enumerados acima serão apresentados utilizando a seguinte notação. Representando-se por Y a variável intervalar resposta e X_j, com $j \in \{1,\dots,p\}$, as variáveis intervalares explicativas, a cada observação i, $i \in \{1,\dots,n\}$ corresponde uma variável resposta intervalar $I_{Y(i)} = [l_{Y(i)}, u_{Y(i)}]$ ou $I_{Y(i)} = [c_{Y(i)} - r_{Y(i)}, c_{Y(i)} + r_{Y(i)}]$ e a cada variável intervalar explicativa corresponde o intervalo $I_{X_{ij}} = [l_{X_{ij}}, u_{X_{ij}}]$ ou $I_{X_{ij}} = [c_{X_{ij}} - r_{X_{ij}}, c_{X_{ij}} + r_{X_{ij}}]$, sendo que $l_{Y(i)}$ e $l_{X_{ij}}$ correspondem aos limites inferiores e $u_{Y(i)}$ e $u_{X_{ij}}$ aos limites superiores dos intervalos $I_{Y(i)}$ e $I_{X_{ij}}$, respetivamente; $c_{Y(i)} = \frac{u_{Y(i)} + l_{Y(i)}}{2}$ e $r_{Y(i)} = \frac{u_{Y(i)} - l_{Y(i)}}{2}$ ao centro e raio da variável resposta e $c_{X_{ij}} = \frac{u_{X_{ij}} + l_{X_{ij}}}{2}$ e $r_{X_{ij}} = \frac{u_{X_{ij}} - l_{X_{ij}}}{2}$ ao centro e raio de cada uma das j variáveis explicativas. No modelo proposto por Dias e Brito (2017) os intervalos são representados através de funções quantil assumindo uma distribuição Uniforme. Assim, considerando a variável intervalar $Y(i)$, a sua função quantil é dada por $\Psi^{-1}_{Y(i)}(t) = c_{Y(i)} + r_{Y(i)}(2t - 1)$, com $t \in [0,1]$. A versatilidade deste modelo permite considerar outras distribuições além da distribuição Uniforme, sendo que a expressão apresentada terá de sofrer alterações.

3.1.1 Método do Centro

O primeiro modelo de regressão linear para variáveis intervalares é designado por Método do Centro e foi proposto por Billard e Diday (2000). Este modelo permite prever uma variável intervalar resposta Y a partir das variáveis intervalares explicativas X_j , com $j \in \{1,\dots,p\}$. Os coeficientes do modelo são estimados aplicando um modelo de regressão linear clássico ao ponto médio dos intervalos, $c_{Y(i)}$ e $c_{X_{ij}}$, com $j \in \{1,\dots,p\}$.

Assim, considerando as variáveis intervalares $X_1, \dots, X_p$ relacionadas com a variável intervalar Y de acordo com uma relação de regressão linear, temos:

$$c_{Y(i)} = \beta_0 + \beta_1 c_{X_{i1}} + \cdots + \beta_p c_{X_{ip}} + \varepsilon_i^c \tag{3.2}$$

Os parâmetros $\beta_0, \beta_1, \dots, \beta_p$ são obtidos através da resolução de um problema de otimização, ou seja, podemos estimar os valores de cada parâmetro β_j, com $j \in \{0, \dots, p\}$, através da minimização da soma dos quadrados dos resíduos (método dos mínimos quadrados) que representam os erros de ajuste do modelo em relação ao centro do intervalo. Neste modelo o problema de otimização é definido por:

$$min\, S = \sum_{i=1}^{n} (\varepsilon_i^c)^2 = \sum_{i=1}^{n} \left(c_{Y(i)} - \beta_o - \beta_1 c_{X_{i1}} - \cdots - \beta_p c_{X_{ip}} \right)^2 \tag{3.3}$$

Os valores dos parâmetros β_j, com $j \in \{0, \dots, p\}$, da expressão (3.2) podem ser calculados diferenciando a expressão (3.3) em relação aos parâmetros β_j, sendo que estes podem assumir qualquer número real.

Neste modelo, os parâmetros da regressão linear são estimados a partir do ponto médio dos intervalos, sendo o modelo posteriormente aplicado aos extremos superiores e inferiores dos intervalos observados para cada X_{ij}, com $j \in \{1, \dots, p\}$, separadamente, de modo a prever os intervalos $Y(i)$.

Como os parâmetros β_j, com $j \in \{0, \dots, p\}$, podem assumir qualquer número real, incluindo números negativos, os intervalos previstos para cada $Y(i)$ podem não ser um intervalo, ou seja, o limite superior previsto para o intervalo pode ser um valor menor do que o previsto para o limite inferior. Assim, na previsão de cada intervalo $Y(i)$ é necessário escolher para o limite inferior o menor valor obtido e para o limite superior o maior valor. Ou seja, as previsões para os limites da variável intervalar Y são obtidos da seguinte forma:

$$l_{\hat{Y}(i)} = min \left\{ \beta_0 + \beta_1 l_{X_{i1}} + \cdots + \beta_p l_{X_{ip}}; \beta_0 + \beta_1 u_{X_{i1}} + \cdots + \beta_p u_{X_{ip}} \right\} \tag{3.4}$$

$$u_{\hat{Y}(i)} = max \left\{ \beta_0 + \beta_1 l_{X_{i1}} + \cdots + \beta_p l_{X_{ip}}; \beta_0 + \beta_1 u_{X_{i1}} + \cdots + \beta_p u_{X_{ip}} \right\} \tag{3.5}$$

Além da limitação que surge quando os parâmetros β_j, com $j = \{0, \ldots, p\}$, são negativos, este modelo é aplicado apenas a um único ponto de referência, os pontos médios dos intervalos observados. Consequentemente, o método reduz os intervalos a um único valor e como tal esta abordagem não é muito diferente da clássica, em que podemos resumir os dados a uma mediana ou média do "grupo" de observações, sendo que desta forma estamos a perder a variabilidade de informação associada a cada observação.

3.1.2 Método do Mínimo e Máximo

Em 2002, Billard e Diday propuseram um modelo similar ao anterior, denominado método do Mínimo e Máximo. Neste método os extremos previstos para os intervalos resposta são obtidos separadamente, mas neste caso, os coeficientes do modelo são estimados aplicando um modelo de regressão linear clássico aos extremos superiores e outro aos extremos inferiores dos intervalos, i.e., os coeficientes que permitem obter a previsão para os extremos dos intervalos $Y(i)$ não são os mesmos (Billard e Diday 2002).

Para este modelo, a variável resposta Y está relacionada com as variáveis explicativas X_j, com $j \in \{1, \ldots, p\}$, de acordo com as seguintes regressões lineares:

$$l_{Y(i)} = \beta_0^l + \beta_1^l l_{X_{i1}} + \cdots + \beta_p^l l_{X_{ip}} + \varepsilon_i^l \tag{3.6}$$

$$u_{Y(i)} = \beta_0^u + \beta_1^u u_{X_{i1}} + \cdots + \beta_p^u u_{X_{ip}} + \varepsilon_i^u \tag{3.7}$$

Das equações (3.6) e (3.7) é possível determinar a soma dos quadrados dos resíduos para o extremo superior e inferior do intervalo, sendo o problema de otimização que permite obter os parâmetros do modelo representado pela equação:

$$min\, S = \sum_{i=1}^{n} \left(\varepsilon_i^l\right)^2 + \sum_{i=1}^{n} (\varepsilon_i^u)^2 \tag{3.8}$$

Ao contrário do método proposto anteriormente, os parâmetros β_j, com $j = \{0, \ldots, p\}$, para cada um dos limites do intervalo são calculados independentemente. Assim, é possível calcular os valores dos parâmetros β_j^u e β_j^l , com $j = \{0, \ldots, p\}$, que

minimizam a soma dos quadrados dos resíduos, ou seja, diferenciando a expressão (3.8) em ordem aos parâmetros. Como os parâmetros para cada um dos limites do intervalo serão diferentes, as estimativas para o extremo superior e o extremo inferior do intervalo serão obtidas pelas seguintes expressões:

$$l_{\hat{Y}(i)} = \beta_0^l + \beta_1^l l_{X_{i1}} + \cdots + \beta_p^l l_{X_{ip}} \tag{3.9}$$

$$u_{\hat{Y}(i)} = \beta_0^u + \beta_1^u u_{X_{i1}} + \cdots + \beta_p^u u_{X_{ip}} \tag{3.10}$$

Relativamente ao modelo do Centro, este modelo permitiu captar maior variabilidade dos dados, porque são utilizados dois pontos de referência do intervalo em alternativa ao ponto médio. No entanto, um dos problemas existentes no primeiro modelo continua a persistir: o valor previsto para o extremo superior do intervalo poderá ser menor que o valor previsto para o extremo inferior, caso os parâmetros β_j^u ou β_j^l, com $j \in \{0, \ldots, p\}$, sejam negativos, e neste caso não será obtido um intervalo. Assim, é necessário garantir nos extremos previstos para cada um dos intervalos, o valor maior representará o extremo superior e o valor menor o extremo inferior, tal como no modelo anterior.

3.1.3 Método do Centro e Raio

No método do Centro a estimativa dos parâmetros β_j , com $j \in \{0, \ldots, p\}$, era baseada apenas no ponto médio dos intervalos. No entanto, como já foi referido, este método poderá não ser o mais eficaz. Na sequência do método do Centro e com o objetivo de captar uma maior variabilidade inerente aos intervalos, foi proposto por Lima Neto e De Carvalho um novo modelo de regressão linear, que incluída a informação dos centros e dos raios dos intervalos. Este modelo apresentou, face aos anteriores, um desempenho melhor (Lima Neto e De Carvalho 2008).

O Método do Centro e Raio estima os parâmetros β_j, com $j = \{0, \ldots, p\}$, resultantes da relação linear entre a variável intervalar resposta Y e as variáveis intervalares explicativas X_j , com $j \in \{1, \ldots, p\}$, usando a informação dos centros e dos raios dos intervalos observados. (As notações referentes ao centro e raio de um intervalo

foram apresentadas no início da secção 3.1.)

A relação entre os centros e a relação entre os raios de uma variável intervalar resposta Y e os respetivos centros e raios das variáveis intervalares explicativas X_j, com $j \in \{1, \ldots, p\}$, são dadas pelas seguintes equações:

$$c_{Y(i)} = \beta_0^c + \beta_1^c c_{X_{i1}} + \cdots + \beta_p^c c_{X_{ip}} + \varepsilon_i^c \tag{3.11}$$

$$r_{Y(i)} = \beta_0^r + \beta_1^r r_{X_{i1}} + \cdots + \beta_p^r r_{X_{ip}} + \varepsilon_i^r \tag{3.12}$$

Neste método é assumido que os valores dos centros e dos raios dos intervalos são valores independentes.

No método do centro e raio, o problema de otimização que permite obter os valores para os parâmetros do modelo é dado pela expressão:

$$min\ S = \sum_{i=1}^{n} ((\varepsilon_i^c)^2 + (\varepsilon_i^r)^2) \tag{3.13}$$

Os parâmetros β_j^c e β_j^r, com $j \in \{0, \ldots, p\}$, são estimados através da aplicação do método dos mínimos quadrados aos modelos de regressão linear referentes aos centros e aos raios, respetivamente. Na equação (3.13) já é apresentado o problema de minimização simplificado. Com os valores dos parâmetros β_j^c e β_j^r , com $j \in \{0, \ldots, p\}$, é possível prever os valores dos centros e raios da variável intervalar resposta Y através das seguintes expressões:

$$c_{\hat{Y}(i)} = \beta_0^c + \beta_1^c c_{X_{i1}} + \cdots + \beta_p^c c_{X_{ip}} \tag{3.14}$$

$$r_{\hat{Y}(i)} = \beta_0^r + \beta_1^r r_{X_{i1}} + \cdots + \beta_p^r r_{X_{ip}} \tag{3.15}$$

Os intervalos previstos são então obtidos da seguinte forma:

$$I_{\hat{Y}(i)} = \left[c_{\hat{Y}(i)} - r_{\hat{Y}(i)} ; c_{\hat{Y}(i)} + r_{\hat{Y}(i)} \right] \tag{3.16}$$

Os parâmetros β_j^c, com $j \in \{0, \ldots, p\}$, podem assumir qualquer valor real. No entanto, se existirem parâmetros β_j^r, com $j \in \{0, \ldots, p\}$, negativos, os resultados

previstos para os raios das observações de Y podem ser negativos. Neste caso, obteríamos $I_{\hat{Y}(i)}$ que não seriam intervalos.

Surge assim um problema idêntico ao ocorrido nos modelos anteriores, visto que os resultados obtidos para $I_{\hat{Y}(i)}$ poderão não ser intervalos de números reais.

3.1.4 Método do Centro e Raio com Restrições

Os métodos anteriormente propostos apresentavam um problema comum. Se determinados parâmetros do modelo fossem negativos não seria possível prever um intervalo, porque o valor previsto para o extremo superior do intervalo assumia um valor menor do que o respetivo extremo inferior. Lima Neto e De Carvalho com o método do Centro e Raio, mostraram a importância de incluir no modelo de regressão linear a informação referente aos centros e aos raios dos intervalos. Deste modo conseguiram melhorar a qualidade das estimativas e captar a variabilidade dos dados. No entanto, este modelo continua a não assegurar que o resultado final seja um intervalo (Lima Neto e De Carvalho 2008).

Assim, considerando novamente (3.11) e (3.12) como as expressões que definem a relação linear entre os centros e raios da variável intervalar resposta, Y_c e Y_r, e os centros e raios das variáveis intervalares explicativas, X_j^c e X_j^r, com $j \in \{1, \ldots, p\}$, Lima Neto e De Carvalho propuseram a inclusão de uma restrição no modelo de regressão dos raios. Esta restrição obriga a que os parâmetros associados aos raios dos intervalos tenham que ser não negativos, para todas as variáveis, i.e., $\beta_j^r \geq 0$, para $j \in \{0, \ldots, p\}$. Desta forma, o modelo de regressão linear dos raios irá assegurar que $r_{\hat{Y}(i)} \geq 0$, o que significa que o valor previsto para extremo inferior do intervalo será sempre menor ou igual que o previsto para extremo superior (Lima Neto e De Carvalho 2010). A restrição de não negatividade não é aplicada aos parâmetros β_j^c, porque os valores previstos para os centros dos intervalos podem ser um qualquer número real.

Neste modelo, o problema de otimização, que inclui a restrição de não negatividade nos parâmetros dos raios, é o seguinte:

$$min\ S = \sum_{i=1}^{n} (\varepsilon_i^c)^2 + \sum_{i=1}^{n} (\varepsilon_i^r)^2 \tag{3.17}$$

com $\beta_j^r \geq 0, j \in \{0, \ldots, p\}$.

O cálculo dos parâmetros β_j^c, com $j \in \{0, \ldots, p\}$, do modelo de regressão linear é efetuado através da resolução do sistema que resulta da expressão (3.17) em ordem aos vários parâmetros. No entanto, para estimar os valores dos parâmetros β_j^r , com $j \in \{0, \ldots, p\}$, com a restrição de não negatividade, os autores recorreram ao algoritmo de *Lawson and Hanson* adaptado a este modelo de regressão linear (Lima Neto e De Carvalho 2010).

O problema da previsão de Y não ser um intervalo fica solucionada com este método. Os valores obtidos para os raios são não negativos e, consequentemente, o extremo inferior do intervalo irá ser sempre menor ou igual que o respetivo extremo superior. Os valores previstos para os centros e raios dos intervalos são calculados através das equações (3.14) e (3.15), como no método anterior.

No entanto, devido à restrição considerada neste método está a ser imposto que a relação entre os raios fique condicionada, ou seja, a relação entre os raios das variáveis explicativas e a variável resposta torna-se desta forma sempre direta. Esta restrição poderá fazer com que este modelo não capte toda a variabilidade nos dados por restringir o conjunto de soluções. Consequentemente, as estimativas não são necessariamente melhores do que as obtidas com o modelo dos centros e raios.

3.1.5 *Lasso IR-Method*

No modelo anterior, de forma a garantir que o valor previsto para o extremo superior dos intervalos seja sempre maior ou igual ao do extremo inferior, Lima Neto e De Carvalho propuseram o método do Centro e Raio com Restrições, que impõe uma restrição de não negatividade nos parâmetros β_j^r, com $j = \{0, \ldots, p\}$, no modelo de regressão linear entre os raios.

Giordani propôs um modelo de regressão linear, *Lasso-IR*, que permite a existência de parâmetros negativos associados aos raios dos intervalos, mas impõe uma restrição de não negatividade para os raios previstos (Giordani 2014).

Este modelo também aplica dois modelos de regressão linear, um aos centros e outro aos raios dos intervalos. No entanto, neste modelo os parâmetros β_j^c e β_j^r, com $j = \{0, \dots, p\}$, estão relacionados, de forma a que os parâmetros dos centros e dos raios sejam o mais semelhantes possível, o que não acontecia nos métodos anteriores.

Sendo Y a variável intervalar resposta e $X_1, \dots, X_p$ as variáveis intervalares explicativas, a relação linear dos centros e raios é dada por:

$$c_{Y(i)} = \beta_0^c + \beta_1^c c_{X_{i1}} + \cdots + \beta_p^c c_{X_{ip}} + \varepsilon_i^c \tag{3.18}$$

$$\begin{aligned} r_{Y(i)} &= \beta_0^r + \beta_1^r r_{X_{i1}} + \cdots + \beta_p^r r_{X_{ip}} + \varepsilon_i^r = \\ &= (\beta_0^c + \beta_0^a) + (\beta_1^c + \beta_1^a) r_{X_{i1}} \dots + (\beta_p^c + \beta_p^a) r_{X_{ip}} + \varepsilon_i^r \end{aligned} \tag{3.19}$$

Sendo $\beta_j^r = \beta_j^c + \beta_j^a$, com $j \in \{0, \dots, p\}$, β_j^a é o valor que indica quanto o coeficiente do centro difere do coeficiente do raio de uma mesma variável. Para estimar os parâmetros do modelo, o autor minimizou a distância entre os valores observados e previstos, utilizando a distância de *Bertoluzza*.

Para dois intervalos U e V, a distância de *Bertoluzza* define-se da seguinte forma:

$$d_\theta^2 = (U^c - V^c)^2 - \theta(U^r - V^r)^2 \tag{3.20}$$

com $\theta \in [0, 1]$.

A escolha do valor para θ depende da importância relativa dos raios com respeito aos centros. *Giordani* assumiu que $\theta = \frac{1}{3}$ é uma escolha razoável.

Os parâmetros do modelo *Lasso-IR* são então calculados minimizando a soma dos quadrados dos resíduos dada pela expressão (Giordani 2014):

$$min\, S = \| \sum_{i=1}^n \varepsilon_i^c \|^2 + \theta \| \sum_{i=1}^n \varepsilon_i^r \|^2 \tag{3.21}$$

com $\sum_{j=1}^p (\beta_j^c + \beta_j^a) r_{X_{ij}} \geq 0$, onde $i \in \{1, \dots, n\}$ e $j \in \{1, \dots, p\}$.

Neste método é imposta uma restrição aos raios previstos, e não aos parâmetros do modelo de regressão linear dos raios. Assim, a garantia que cada raio previsto não é negativo é dada pela restrição: $\sum_{j=1}^{p}(\beta_j^c + \beta_j^a)r_{X_{ij}} \geq 0$, com $i \in \{1,\ldots,n\}$. Cada variável intervalar $\hat{Y}(i)$ é calculada pela equação (3.16), utilizando os valores previstos para o centro e o raio obtidos pelo método do *Lasso IR*. Apesar de neste modelo continuarem a existir restrições, estas possibilitam que o algoritmo de otimização procure a solução ótima num maior espaço de valores, porque os parâmetros do modelo dos raios podem ser negativos.

3.1.6 *Interval Distribution (ID) Regression Model*

Ao contrário dos modelos anteriormente propostos para variáveis intervalares, o modelo proposto por Brito e Dias em 2017 não utiliza pontos de referência do intervalo, aos quais se aplica o modelo clássico de regressão linear. O *ID Regression Model* utiliza funções quantil para a representação dos intervalos, ou seja, considera todo o intervalo e uma distribuição inerente ao mesmo. As representações dos intervalos por funções quantil foram apresentadas no início da secção 3.1.

Este modelo estima uma variável intervalar resposta Y a partir das variáveis explicativas X_j , com $j \in \{1, \ldots, p\}$, sem a necessidade de decompor os intervalos em centros e raios ou extremos, como nos anteriores modelos propostos (Billard e Diday 2000; Billard e Diday 2002; Lima Neto e De Carvalho 2008; Lima Neto e De Carvalho 2010; Giordani 2014).

Em Dias e Brito (2017) é assumida uma distribuição Uniforme nos intervalos, no entanto a versatilidade deste modelo permite considerar outras distribuições. Ao assumir a distribuição Uniforme num intervalo a função quantil que o representa é uma função linear não decrescente, de domínio [0,1].

A relação linear entre funções quantil não se pode resumir a uma simples generalização do modelo clássico, porque os parâmetros estimados poderão ser negativos e, consequentemente, a função prevista para $Y(i)$, não será uma função não decrescente (característica obrigatória das funções quantil). Foi assim necessário impor uma restrição

de não negatividade nos parâmetros, para garantir que a função prevista fosse sempre uma função quantil. Consequentemente, foi necessário incluir no modelo de regressão linear não só a função quantil que representa cada variável intervalar X_j, mas também a respetiva função quantil simétrica, de modo a garantir que a relação entre as variáveis intervalares pudesse ser direta ou inversa.

Considerado um conjunto de variáveis intervalares X_j, com $j \in \{1,\dots,p\}$, as funções quantil que representam os valores destas variáveis são representadas por $\Psi_{\mathrm{X_{i1}}}^{-1}(\mathrm{t}),\dots,\Psi_{\mathrm{X_{ip}}}^{-1}(\mathrm{t})$ e as funções quantil que representam os intervalos simétricos são representadas por $-\Psi_{\mathrm{X_{i1}}}^{-1}(1-\mathrm{t}),\dots,-\Psi_{\mathrm{X_{ip}}}^{-1}(1-\mathrm{t})$, com $\mathrm{t} \in [0,1]$. Para cada variável resposta Y, cada função quantil $\Psi_{\mathrm{Y(i)}}^{-1}(\mathrm{t})$ pode ser expressa por $\Psi_{\mathrm{Y(i)}}^{-1}(\mathrm{t}) = \Psi_{\widehat{\mathrm{Y}}\mathrm{(i)}}^{-1} + \varepsilon_{\mathrm{i}}(\mathrm{t})$, onde $\Psi_{\widehat{\mathrm{Y}}\mathrm{(i)}}^{-1}(\mathrm{t})$ é a função quantil prevista para cada observação i, e obtida pela seguinte relação linear entre a variável resposta e as variáveis explicativas:

$$\Psi_{\widehat{Y}(i)}^{-1}(t) = \gamma + \sum_{j=1}^{p} \alpha_j \Psi_{X_{ij}}^{-1}(t) - \sum_{j=1}^{p} \beta_j \Psi_{X_{ij}}^{-1}(1-t) \qquad (3.22)$$

com $t \in [0,1];\ \alpha_j, \beta_j \geq 0, j \in \{1,2,\dots,p\}$ e o termo independente $\gamma \in \mathbb{R}$.

Considera-se que as variáveis X_j e Y têm uma relação linear direta quando $\alpha_j > \beta_j$ e que a relação entre as duas variáveis é inversa quando $\alpha_j < \beta_j$.

No modelo proposto, foi assumida uma distribuição uniforme inerente aos intervalos observados, pelo que $\Psi_{Y(i)}^{-1}(t) = c_{X_{ij}} + (2t-1)r_{X_{ij}}$ e $-\Psi_{Y(i)}^{-1}(1-t) = -c_{X_{ij}} + (2t-1)r_{X_{ij}}$, sendo que a função quantil prevista (3.22) poderá ser escrita da seguinte forma:

$$\Psi_{\widehat{Y}(i)}^{-1}(t) = \sum_{j=1}^{p} (\alpha_j - \beta_j) c_{X_{ij}} + \gamma + \sum_{j=1}^{p} (\alpha_j + \beta_j) r_{X_{ij}} (2t-1) \qquad (3.23)$$

com $t \in [0,1];\ \alpha_j, \beta_j \geq 0, j \in \{1,2,\dots,p\}$ e $\gamma \in \mathbb{R}$.

Na forma de intervalo, cada $\widehat{Y}_i$ é obtida pela expressão (3.23), sendo os extremos

inferiores e superiores determinados da seguinte forma:

$$\Psi_{\hat{Y}(i)}^{-1}(0) = \sum_{j=1}^{p} \alpha_j \left(c_{X_{ij}} - r_{X_{ij}}\right) - \sum_{j=1}^{p} \beta_j \left(c_{X_{ij}} + r_{X_{ij}}\right) + \gamma \tag{3.24}$$

$$\Psi_{\hat{Y}(i)}^{-1}(1) = \sum_{j=1}^{p} \alpha_j \left(c_{X_{ij}} + r_{X_{ij}}\right) - \sum_{j=1}^{p} \beta_j \left(c_{X_{ij}} - r_{X_{ij}}\right) + \gamma \tag{3.25}$$

sendo $I_{\hat{Y}(i)} = [\Psi_{Y(i)}^{-1}(0); \Psi_{Y(i)}^{-1}(1)]$.

Consequentemente, o centro e o raio de um intervalo estimado, $I_{\hat{Y}(i)}$, podem ser obtidos através das seguintes relações lineares:

$$c_{\hat{Y}(i)} = \sum_{j=1}^{p} (\alpha_j - \beta_j) c_{X_{ij}} + \gamma \tag{3.26}$$

$$r_{\hat{Y}(i)} = \sum_{j=1}^{p} (\alpha + \beta_j) r_{X_{ij}} \tag{3.27}$$

Os parâmetros do modelo são estimados através da resolução de um problema de otimização de mínimos quadrados, sujeito a restrições de não negatividade nos parâmetros. Neste problema é utilizada a distância de *Mallows* que considerando a distribuição Uniforme inerente aos intervalos pode ser definida da seguinte forma:

$$D_M^2\left(\Psi_{Y(i)}^{-1}, \Psi_{X(i)}^{-1}\right) = \int_0^1 \left(\Psi_{Y(i)}^{-1}(t) - \Psi_{X(i)}^{-1}(t)\right)^2 dt = \left(c_{Y(i)} - c_{X_j(i)}\right)^2 + \frac{1}{3}\left(r_{Y(i)} - r_{X_j(i)}\right)^2 \tag{3.28}$$

Assim, o problema de otimização que nos permite obter os valores para os parâmetros do *ID Model* é o seguinte:

$$min \sum_{i=1}^{n} \left[\left(c_{Y(i)} - \sum_{j=1}^{p} (\alpha_j - \beta_j) c_{X_{ij}} - \gamma \right)^2 + \frac{1}{3} \left(r_{Y(i)} - \sum_{j=1}^{p} (\alpha + \beta_j) r_{X_{ij}} \right)^2 \right] \tag{3.29}$$

com $t \in [0,1]$; $\alpha_j, \beta_j \geq 0, j \in \{1, 2, \ldots, p\}$ e $\gamma \in \mathbb{R}$.

Este modelo relaciona os intervalos associados às variáveis resposta e explicativas como um todo. Desde modo, a potencialidade do modelo em captar a variabilidade inerente aos dados é superior à dos restantes modelos, onde apenas são considerados os pontos de referência do intervalo, tais como o centro e raio ou extremos.

Este modelo será o ponto de partida para o desenvolvimento desta dissertação, considerando a versatilidade do método para assumir outra distribuição no intervalo, além da distribuição Uniforme. Dado que todos os conceitos e métodos de análise foram desenvolvidos sob o pressuposto da distribuição Uniforme, quando consideramos outra distribuição será necessário rever e desenvolver novos conceitos e métodos.

Capítulo 4

ID Regression Model: Novas Abordagens

Neste capítulo será realizado um estudo mais detalhado do *ID Regression Model*, apresentado no capítulo anterior, quando é assumida uma distribuição Uniforme, inerente aos intervalos observados, mas estendendo o modelo, considerando agora também uma distribuição Triangular, Simétrica ou Geral, em cada intervalo. Nas primeiras secções deste capítulo, serão apresentadas as definições de funções quantil e as expressões para a distância de *Mallows*, com um maior detalhe, para cada uma das distribuições consideradas neste estudo. Tendo em conta a versatilidade do modelo em considerar outras distribuições no intervalo, Dias e Brito (2017) apresentaram uma extensão do modelo, quando é assumida a distribuição Triangular Simétrica nos Intervalos. No seguimento do trabalho destas autoras, será proposta nesta dissertação uma extensão do *ID Model* ao caso em que a distribuição Triangular Geral é considerada.

4.1 *ID Model* e a representação dos intervalos por funções quantil

Como já foi referido no capítulo anterior, neste modelo as variáveis intervalares são representadas por funções quantil e não apenas por pontos de referência, como os extremos ou centros e raios dos intervalos. A utilização de funções quantil para representar histogramas foi proposta por Irpino e Verde (2006), sendo que Bertoluzza, Corral, e Salas (1995) já tinham utilizado uma representação semelhante para intervalos denominada por "parametrização do intervalo". A representação de intervalos por funções quantil, no âmbito da Análise Simbólica de Dados, surge em Dias (2014).

Uma função quantil é a função inversa da função distribuição acumulada. Esta é uma função de domínio $[0,1]$ sempre não decrescente. Uma função quantil para uma variável intervalar Y será representada por $\Psi_{Y(i)}^{-1}(t)$, com $t \in [0,1]$.

Se considerássemos um modelo de regressão linear semelhante ao clássico impondo apenas restrições de não negatividade aos parâmetros, estaríamos a forçar uma

relação linear direta entre as variáveis. Para ultrapassar o problema, Dias e Brito (2017) propuseram o *ID Model*. Este modelo é definido à custa das funções quantil que representam as variáveis intervalares explicativas $\Psi_{X_j(i)}^{-1}(t)$, e as funções quantil que representam os respetivos intervalos simétricos, $-\Psi_{X_j(i)}^{-1}(1-t)$. Deste modo, a relação linear entre os intervalos não será necessariamente direta, apesar das restrições de não negatividade impostas aos parâmetros. Neste modelo, a relação linear entre a variável resposta $Y(i)$ e as variáveis explicativas X_{ij}, com $j \in \{1, \ldots, p\}$ é dada por:

$$\Psi_{Y(i)}^{-1}(t) = \sum_{j=1}^{p} \alpha_j \Psi_{X_{ij}}^{-1}(t) - \sum_{j=1}^{p} \beta_j \Psi_{X_{ij}}^{-1}(1-t) + \gamma \tag{4.1}$$

A função quantil associada a um intervalo e a função quantil associada ao intervalo simétrico têm estruturas diferentes dependendo da distribuição inerente a cada intervalo. Nesta secção é apresentada a definição de uma função quantil quando uma distribuição Uniforme, uma distribuição Triangular Simétrica ou uma distribuição Triangular Geral são assumidas nos intervalos.

Distribuição Uniforme

No modelo inicial proposto por Dias (2014), as funções quantil são representadas assumindo uma distribuição Uniforme nos intervalos. Cada intervalo Y pode, usando o seu centro e raio, ser representado pela seguinte função quantil:

$$\Psi_{Y(i)}^{-1}(t) = c_{Y(i)} + r_{Y(i)}(2t-1), \quad 0 \leq t \leq 1 \tag{4.2}$$

A função quantil que representa o intervalo simétrico é representada da seguinte forma:

$$-\Psi_{Y(i)}^{-1}(1-t) = -c_{Y(i)} + r_{Y(i)}(2t-1), \quad 0 \leq t \leq 1 \tag{4.3}$$

Exemplo: Considerando o intervalo $I_Y = [1,4]$ e assumindo a distribuição Uniforme, o intervalo é representado pela uma função quantil, definida à custa do seu centro e raio, através da seguinte expressão:

$$\Psi_{I_Y}^{-1}(t) = \frac{5}{2} + \frac{3}{2}(2t-1), \quad 0 \leq t \leq 1 \tag{4.4}$$

O intervalo simétrico $-I_Y = [-4, -1]$ pode ser representado através da seguinte expressão:

$$-\Psi_{I_Y}^{-1}(1-t) = -\frac{5}{2} + \frac{3}{2}(2t-1), \quad 0 \leq t \leq 1 \tag{4.5}$$

Estes intervalos e as respetivas funções quantil são representadas graficamente da forma que se segue, estando à esquerda representados os intervalos e à direita as respetivas funções quantil.

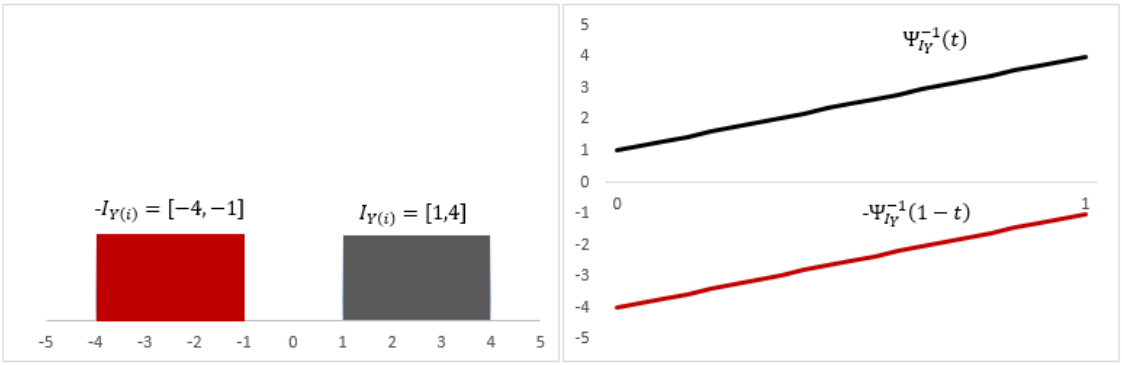


Figura 4.1: Representação de um intervalo, do seu intervalo simétrico e respetivas funções quantil, assumindo uma distribuição Uniforme em cada intervalo

Quando é considerada uma distribuição Uniforme no intervalo, é assumido que todos os valores inerentes ao mesmo crescem uniformemente desde o extremo inferior ao extremo superior do intervalo, como é possível verificar através das funções quantil representadas na Figura (4.1).

Distribuição Triangular Simétrica

Como já foi mencionado nesta dissertação, Dias e Brito (2017) demonstraram a versatilidade deste modelo para assumir outras distribuições inerentes às variáveis intervalares ao considerarem uma distribuição Triangular Simétrica.

Assim, assumindo uma distribuição Triangular Simétrica em cada intervalo $Y(i)$, com centro $c_{Y(i)}$ e raio $r_{Y(i)}$, o intervalo $Y(i)$ pode ser representado por uma função quantil, da seguinte forma:

$$\Psi_{Y(i)}^{-1}(t) = \begin{cases} c_{Y(i)} - r_{Y(i)} + r_{Y(i)}\sqrt{2t}, & 0 \leq t < \frac{1}{2} \\ c_{Y(i)} + r_{Y(i)} - r_{Y(i)}\sqrt{2(1-t)}, & \frac{1}{2} \leq t < 1 \end{cases} \quad (4.6)$$

Neste caso, a função quantil que representa o intervalo é uma função definida por dois ramos, de domínio $[0, 1]$, com a mudança de ramo em $t = \frac{1}{2}$.

A função quantil que representa o intervalo simétrico é representada da seguinte forma:

$$-\Psi_{Y(i)}^{-1}(1-t) = \begin{cases} -c_{Y(i)} - r_{Y(i)} + r_{Y(i)}\sqrt{2t}, & 0 \leq t < \frac{1}{2} \\ -c_{Y(i)} + r_{Y(i)} - r_{Y(i)}\sqrt{2(1-t)}, & \frac{1}{2} \leq t < 1 \end{cases} \quad (4.7)$$

Exemplo: Considerando o intervalo $I_Y = [1,4]$ e assumindo uma distribuição Triangular Simétrica, o intervalo, definido à custa do seu centro e raio, é representado pela função quantil que se segue:

$$\Psi_{I_Y}^{-1}(t) = \begin{cases} 1 + \frac{3}{2}\sqrt{2t}, & 0 \leq t < \frac{1}{2} \\ 4 - \frac{3}{2}\sqrt{2(1-t)}, & \frac{1}{2} \leq t < 1 \end{cases} \quad (4.8)$$

O intervalo simétrico $-I_Y = [-4, -1]$ pode ser representado através da seguinte expressão:

$$-\Psi_{I_Y}^{-1}(1-t) = \begin{cases} -4 + \frac{3}{2}\sqrt{2t}, & 0 \leq t < \frac{1}{2} \\ -1 - \frac{3}{2}\sqrt{2(1-t)}, & \frac{1}{2} \leq t < 1 \end{cases} \quad (4.9)$$

Estes intervalos e as respetivas funções quantil são representadas graficamente da forma que se segue, estando à esquerda representados os intervalos e à direita as respetivas funções quantil.

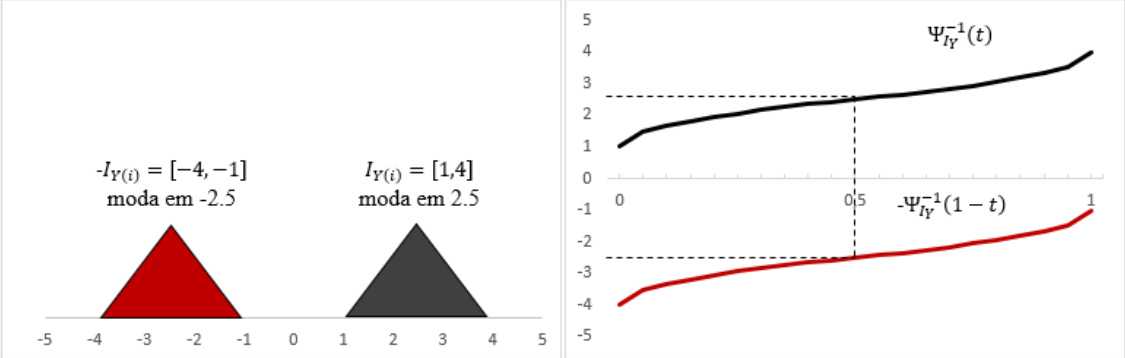


Figura 4.2: Representação de um intervalo, do seu intervalo simétrico e respetivas funções quantil, assumindo uma distribuição Triangular Simétrica em cada intervalo

Quando é considerada uma distribuição Triangular Simétrica nos dados, os mesmos são estruturados tendo em consideração a Moda do intervalo, ou seja, o ponto de inflexão da função quantil. Este desenvolvimento vai enriquecer o modelo, na medida em que contempla mais informação de cada uma das observações das variáveis.

Distribuição Triangular Geral

Nesta dissertação será proposto a extensão do *ID Model* para variáveis intervalares que assumam uma distribuição Triangular Geral. Assim, foi necessário definir a função quantil que representa os intervalos com essa distribuição. Neste caso, é também necessário considerar mais um ponto de referência do intervalo, a Moda, que pode ser qualquer valor contido no intervalo.

A distribuição Triangular é uma distribuição contínua definida num intervalo que é descrita pelo seu valor mínimo, o seu valor máximo e a sua moda ou pelo o centro, o raio e a moda. A função densidade associada poderá ser simétrica, como no caso apresentado na secção anterior em que a moda do intervalo é igual ao seu centro, ou assimétrica, quando a moda não coincide com o centro do intervalo.

Assumindo uma distribuição Triangular Geral, Cheira et al (2017) representa cada intervalo $Y(i)$, com extremo inferior $l_{Y(i)}$, extremo superior $u_{Y(i)}$ e moda $m_{Y(i)}$ por uma função quantil da seguinte forma:

$$\Psi_{Y(i)}^{-1}(t) = \begin{cases} l_{Y(i)} + \sqrt{\left(u_{Y(i)} - l_{Y(i)}\right)\left(m_{Y(i)} - l_{Y(i)}\right)t}, & 0 \le t \le \dfrac{m_{Y(i)} - l_{Y(i)}}{u_{Y(i)} - l_{Y(i)}} \\ u_{Y(i)} - \sqrt{\left(u_{Y(i)} - l_{Y(i)}\right)\left(u_{Y(i)} - m_{Y(i)}\right)(1-t)}, & \dfrac{m_{Y(i)} - l_{Y(i)}}{u_{Y(i)} - l_{Y(i)}} < t \le 1 \end{cases} \quad (4.10)$$

A função quantil que representa o intervalo $Y(i)$ é uma função de domínio [0,1], definida por dois ramos, com a mudança de ramo em $t = \frac{m_{Y(i)} - l_{Y(i)}}{u_{Y(i)} - l_{Y(i)}}$.

A função quantil que representa o intervalo simétrico é representada da seguinte forma:

$$-\Psi_{Y(i)}^{-1}(1-t) = \begin{cases} -u_{Y(i)} + \sqrt{\left(u_{Y(i)} - l_{Y(i)}\right)\left(u_{Y(i)} - m_{Y(i)}\right)t}, & 0 \le t \le \dfrac{u_{Y(i)} - m_{Y(i)}}{u_{Y(i)} - l_{Y(i)}} \\ -l_{Y(i)} - \sqrt{\left(u_{Y(i)} - l_{Y(i)}\right)\left(m_{Y(i)} - l_{Y(i)}\right)(1-t)}, & \dfrac{u_{Y(i)} - m_{Y(i)}}{u_{Y(i)} - l_{Y(i)}} < t \le 1 \end{cases} \quad (4.11)$$

A mudança de ramo desta função é em $t = \frac{u_{Y(i)} - m_{Y(i)}}{u_{Y(i)} - l_{Y(i)}}$, não coincidindo com a mudança de ramo da função quantil em (4.10).

Exemplo: Considerando o intervalo $I_Y = [1,4]$ com $m_{I_Y} = 3$ e assumindo uma distribuição Triangular Geral, o intervalo é representado por uma função quantil através da seguinte expressão:

$$\Psi_{I_Y}^{-1}(t) = \begin{cases} 1 + \sqrt{6t}, & 0 \le t \le \dfrac{2}{3} \\ 4 - \sqrt{3(1-t)}, & \dfrac{2}{3} < t \le 1 \end{cases} \quad (4.12)$$

O intervalo simétrico $-I_Y = [-4,-1]$, com $m_{-I_Y} = -3$ pode ser representado por:

$$-\Psi_{I_Y}^{-1}(1-t) = \begin{cases} -4 + \sqrt{3t}, & 0 \le t \le \dfrac{1}{3} \\ -1 - \sqrt{6(1-t)}, & \dfrac{1}{3} < t \le 1 \end{cases} \quad (4.13)$$

Estes intervalos e as respetivas funções quantil são representadas graficamente da forma que se segue, estando à esquerda representados os intervalos e à direita as respetivas funções quantil.

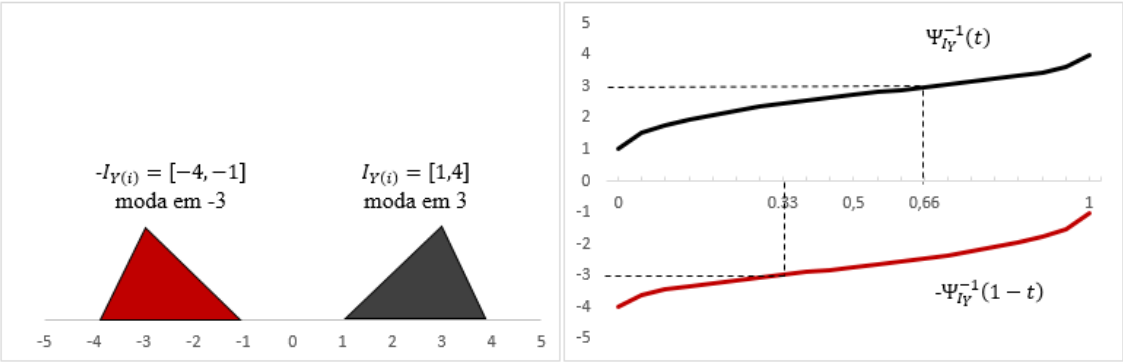


Figura 4.3: Representação de um intervalo, do seu intervalo simétrico e respetivas funções quantil, assumindo uma distribuição Triangular Geral em cada intervalo

Neste exemplo é possível observar que a moda do intervalo e a moda do seu intervalo simétrico não são simétricas, ao contrario do que acontece no caso de uma distribuição Triangular Simétrica. Esta diferença no valor das modas irá dificultar a combinação linear do modelo, porque a relação linear entre a função quantil que representa um intervalo e a função quantil que representa o seu simétrico, será uma função definida por três ramos.

4.2 Distância de *Mallows*

De forma estimar os valores dos parâmetros e a avaliar o desempenho do modelo será necessário medir a semelhança entre os intervalos previstos e observados, para o que é necessário selecionar uma distância que permita cumprir este objetivo. Na literatura várias distâncias foram estudadas (Arroyo e Maté 2009; Irpino e Verde 2015) e concluiu-se que a distância de *Mallows* (Mallows 1972) é considerada uma medida adequada para avaliar a similaridade entre intervalos representados por funções quantil. Esta medida tem interpretações intuitivas relacionadas com a *Earth Mover's Distance* e é a distância que melhor se ajusta ao conceito comum de distância captada pela visão humana (Arroyo e Maté 2009). Esta distância também já foi considerada na análise classificatória e em modelos de regressão linear para variáveis histograma (Irpino e Verde 2006; Irpino e Verde 2015; Dias e Brito 2017) e referenciada no contexto da classificação (Hofer 2015). Assim, a distância que irá ser utilizada para quantificar a semelhança entre intervalos representados através de funções quantil é a distância de *Mallows*.

O quadrado da distância de *Mallows* entre duas funções quantil, $\Psi_X^{-1}(t)$ e $\Psi_Y^{-1}(t)$, que representam os valores das variáveis intervalares X e Y, respetivamente, é definida da seguinte forma (Mallows 1972):

$$D_M^2(\Psi_X^{-1}, \Psi_Y^{-1}) = \int_0^1 \left(\Psi_X^{-1}(t) - \Psi_Y^{-1}(t)\right)^2 dt \tag{4.14}$$

Quando é considerado que os intervalos têm inerente uma distribuição Uniforme ou uma distribuição Triangular Simétrica, a distância reduz-se a uma simples expressão. No entanto, quando é assumida uma distribuição Triangular Geral a expressão da distância torna-se mais complexa devido ao facto de ser necessário introduzir a moda para definir os intervalos.

Irpino e Verde (2006) obtiveram uma expressão simplificada para o quadrado da distância de *Mallows* entre dois histogramas representados por funções quantil, e onde se assume a distribuição Uniforme nos subintervalos. Particularizando essa expressão para o caso dos intervalos com uma distribuição Uniforme, a expressão (4.14) pode ser reescrita como a soma dos quadrados das diferenças entre os centros e um terço dos quadrados das diferenças entre os raios dos intervalos. Dadas duas funções quantil $\Psi_Y^{-1}(t)$ e $\Psi_X^{-1}(t)$, que representam os intervalos Y e X, respetivamente, a expressão do quadrado da distância de *Mallows* definida em (4.14) escreve-se da seguinte forma:

$$D_M^2(\Psi_X^{-1}, \Psi_Y^{-1}) = (c_X - c_Y)^2 + \frac{1}{3}(r_X - r_Y)^2 \tag{4.15}$$

onde c_X e c_Y são os centros e r_X e r_Y os raios dos intervalos de X e Y, respetivamente.

Quando é considerada a distribuição Triangular Simétrica em cada intervalo, Dias e Brito (2017) obtiveram para o quadrado da distância de *Mallows* uma expressão semelhante.

Dadas duas funções quantil $\Psi_X^{-1}(t)$ e $\Psi_Y^{-1}(t)$, que representam os intervalos X e Y, respetivamente, e assumindo uma distribuição Triangular Simétrica, o quadrado da distância de *Mallows* definida em (4.14), pode ser reescrita da seguinte forma:

$$D_M^2(\Psi_X^{-1}, \Psi_Y^{-1}) = (c_X - c_Y)^2 + \frac{1}{6}(r_X - r_Y)^2 \tag{4.16}$$

onde c_X e c_Y são os centros e r_X e r_Y os raios dos intervalos de X e Y, respetivamente.

É de notar que o peso associado à diferença entre os raios é menor quando os intervalos assumem uma distribuição Triangular Simétrica do que quando é assumida a distribuição Uniforme.

A partir da expressão (4.14), Cheira et al (2017) obtiveram uma expressão para o quadrado da distância de *Mallows* no contexto de variáveis intervalares, quando se assume uma distribuição Triangular Geral. Neste caso, a expressão obtida é bastante mais complexa do que as anteriores.

Quando assumimos a distribuição Triangular Geral num intervalo, este já não pode ser representado apenas pelos seus extremos ou centro e raio, é sempre necessário indicar também o valor da moda. Assim, neste caso será considerado que cada intervalo é representado pelo vetor $(l, u, m) = (c - r, c + r, m)$, em que l representa o extremo inferior, u o extremo superior, c o centro, r o raio e m a moda do intervalo.

Quando $X = (c_X, r_X, m_X)$ e $Y = (c_Y, r_Y, m_Y)$, com $r_X \, e \, r_Y > 0$, o quadrado da distância de *Mallows* entre dois intervalos representados pelas funções quantil $\Psi_X^{-1}(t) \, e \, \Psi_Y^{-1}(t)$, assumindo uma distribuição Triangular, é dado por:

- Se $\frac{m_X - c_X}{2r_X} \leq \frac{m_Y - c_Y}{2r_Y}$:

$$
\begin{aligned}
D_m^2 = {} & (c_X - c_Y)^2 + \frac{1}{6}(r_X - r_Y)^2 + \frac{1}{6}(m_X - c_X)^2 + \frac{1}{6}(m_Y - c_Y)^2 - \frac{5}{3} r_X r_Y \\
& + \frac{2}{3}(m_X - c_X)(c_X - c_Y + r_Y) - \frac{2}{3}(m_Y - c_Y)(c_X - c_Y + r_X) \\
& + \frac{\sqrt{r_X r_Y}}{6} \sqrt{m_X - c_X + r_X} \sqrt{m_Y - c_Y + r_Y} \left(5 - \frac{m_X - c_X}{r_X}\right) \\
& + \frac{\sqrt{r_X r_Y}}{6} \sqrt{c_X + r_X - m_X} \sqrt{c_Y + r_Y - m_Y} \left(5 - \frac{m_Y - c_Y}{r_Y}\right) \\
& + \frac{\sqrt{r_X r_Y}}{3} \sqrt{c_X + r_X - m_X} \sqrt{m_Y - c_Y + r_Y} \left(\operatorname{arcsen} \frac{m_Y - c_Y}{r_Y}\right. \\
& \left. - \operatorname{arcsen} \frac{m_X - c_X}{r_X}\right) \left(\operatorname{arcsen} \frac{m_X - c_X}{r_X}\right. \\
& \left. - \operatorname{arcsen} \frac{m_Y - c_Y}{r_Y}\right)
\end{aligned}
\tag{4.17}
$$

- Se $\frac{m_X - c_X}{2r_x} > \frac{m_Y - c_Y}{2r_Y}$:

$$
\begin{aligned}
D_m^2 = {} & (c_X - c_Y)^2 + \frac{1}{6}(r_X - r_Y)^2 + \frac{1}{6}(m_X - c_X)^2 + \frac{1}{6}(m_Y - c_Y)^2 - \frac{5}{3} r_X r_Y \\
& + \frac{2}{3}(m_X - c_X)(c_X - c_Y - r_Y) - \frac{2}{3}(m_Y - c_Y)(c_X - c_Y - r_X) \\
& + \frac{\sqrt{r_X r_Y}}{6} \sqrt{m_X - c_X + r_X} \sqrt{m_Y - c_Y + r_Y} \left(5 - \frac{m_Y - c_Y}{r_Y}\right) \\
& + \frac{\sqrt{r_X r_Y}}{6} \sqrt{c_X + r_X - m_X} \sqrt{c_Y + r_Y - m_Y} \left(5 - \frac{m_X - c_X}{r_X}\right) \\
& + \frac{\sqrt{r_X r_Y}}{3} \sqrt{c_Y + r_Y - m_Y} \sqrt{m_X - c_X + r_X} \left(\operatorname{arcsen} \frac{m_X - c_X}{r_X} \right. \\
& \left. - \operatorname{arcsen} \frac{m_Y - c_Y}{r_Y}\right)
\end{aligned}
\tag{4.18}
$$

Quando $X = (c_X, r_X, m_X)$ e $Y = (c_Y, 0, c_Y)$, com $r_X > 0$ e $r_Y = 0$, o quadrado da distância de *Mallows* entre estes dois intervalos representados pelas funções quantil $\Psi_X^{-1}(t)$ e $\Psi_Y^{-1}(t) = c_Y$, assumindo uma distribuição Triangular, é dado por:

$$
\begin{aligned}
D_m^2 = {} & (c_X - c_Y)^2 - \frac{1}{3} r_X^2 - \frac{4}{3}(m_X - c_X)^2 + \frac{2}{3}(m_X - c_X)(c_X - c_Y) + \frac{(m_X - c_X + r_X)^3}{4 r_X} \\
& + \frac{(c_X + r_X - m_X)^3}{4 r_X}
\end{aligned}
\tag{4.19}
$$

Quando $X = (c_X, 0, c_X)$ e $Y = (c_Y, 0, c_Y)$, com $r_X, r_Y = 0$, estamos no caso particular em que X e Y são dois valores reais. Neste caso o quadrado da distância de *Mallows* é dado por:

$$
D_m^2 = (c_X - c_Y)^2 \tag{4.20}
$$

A partir das expressões obtidas para o quadrado da distância de *Mallows* quando se assume a distribuição Triangular, é possível obter os parâmetros do modelo de regressão e calcular as respetivas medidas de qualidade.

Ao contrário da expressão para o quadrado da distância de *Mallows* quando se assume a distribuição Uniforme ou a distribuição Triangular Simétrica, em que para todas as observações, a distância é calculada da mesma forma, no caso da distribuição Triangular Geral é necessário estabelecer um algoritmo que defina para cada observação, de acordo com a estrutura do intervalo, qual a expressão da distância a utilizar.

4.3 Medidas de Avaliação do Modelo

Segundo Bock e Diday (2000), a média simbólica para variáveis intervalares é dada por $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} c_{Y(i)}$. Utilizando a distância de *Mallows*, Dias e Brito (2017) provaram que quando a distribuição dentro do intervalo é Uniforme ou Triangular Simétrica, a soma do quadrado da distância de *Mallows* entre cada observação $Y(i)$ e a média simbólica da variável $Y(i)$, isto é $\bar{Y}$, pode decompor-se da seguinte forma:

$$\sum_{i=1}^{n} D_M^2\left(\Psi_{Y(i)}^{-1}(t), \bar{Y}\right) = \sum_{i=1}^{n} D_M^2\left(\Psi_{Y(i)}^{-1}(t), \Psi_{\hat{Y}(i)}^{-1}(t)\right) + \sum_{i=1}^{n} D_M^2\left(\Psi_{\hat{Y}(i)}^{-1}(t), \bar{Y}\right) \qquad (4.21)$$

onde $\bar{Y} = \sum_{j=1}^{p}\left(\alpha_j^* - \beta_j^*\right)\bar{X}_j + \gamma^*$, sendo que α^*, β^*e γ^* constituem a solução ótima para os parâmetros do modelo e $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} c_{Y(i)}$ (média simbólica para variáveis intervalares - Bock e Diday 2000).

Esta decomposição permite definir uma medida da qualidade do ajuste para variáveis intervalares do *ID Model* quando a distribuição no intervalo é Uniforme ou Triangular Simétrica. Considerando a função quantil dos valores observados, $\Psi_{Y(i)}^{-1}(t)$, a função quantil prevista $\Psi_{\hat{Y}(i)}^{-1}(t)$, com $t \in [0,1]$, e a média simbólica da variável Y a medida da qualidade do ajuste do modelo, Ω, proposta por Dias e Brito (2017) é dada por:

$$\Omega = \frac{\sum_{i=1}^{n} D_M^2\left(\Psi_{\hat{Y}(i)}^{-1}(t), \bar{Y}\right)}{\sum_{i=1}^{n} D_M^2\left(\Psi_{Y(i)}^{-1}(t), \bar{Y}\right)} \qquad (4.22)$$

Esta medida de avaliação do desempenho do modelo varia entre zero e um.

É de salientar, que não foi demonstrada a decomposição apresentada na expressão (4.22) para o caso da distribuição Triangular Geral. Por este motivo a medida Ω não é aplicável para este caso.

Outras medidas que irão ser utilizadas para avaliar a similaridade entre os intervalos observados e previstos são: a raiz do Erro Quadrático Médio para os extremos inferiores, que mede o ajuste dos extremos inferiores dos intervalos observados e dos intervalos previstos, a raiz do Erro Quadrático Médio para o extremos superiores, que

mede o ajuste dos extremos superiores dos intervalos observados e previstos (Lima Neto e De Carvalho, 2008; Lima Neto e De Carvalho, 2010) e a raiz do Erro Quadrático Médio definido pela distância de *Mallows* (Irpino e Verde 2015):

$$RMSE_L = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(l_{\hat{Y}(i)} - l_{Y(i)}\right)^2} \tag{4.23}$$

$$RMSE_U = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(u_{\hat{Y}(i)} - u_{Y(i)}\right)^2} \tag{4.24}$$

$$RMSE_M = \sqrt{\frac{1}{n}\sum_{i=1}^{n} D_M^2\left(\Psi_{Y(i)}^{-1}(t), \Psi_{\hat{Y}(i)}^{-1}(t)\right)} \tag{4.25}$$

O desempenho dos modelos para cada uma das distribuições consideradas e para os restantes modelos apresentados no Capítulo 3, será avaliado e comparado através destas medidas.

4.4 *ID Regression Model:* Distribuição Uniforme

A relação linear entre uma variável intervalar resposta, Y, e as variáveis explicativas, X_j, com $j \in \{1, \dots, p\}$, de acordo com as funções quantil definidas para uma distribuição Uniforme em (4.2) e (4.3) é dada por:

$$\Psi_{\hat{Y}(i)}^{-1}(t) = \sum_{j=1}^{p}\left(\alpha_j - \beta_j\right)c_{X_{ij}} + \gamma + \sum_{j=1}^{p}\left(\alpha_j + \beta_j\right)r_{X_{ij}}(2t-1) \tag{4.26}$$

com $t \in [0,1]$; $\alpha_j, \beta_j \geq 0, i \in \{1, 2, \dots, p\}$ e $\gamma \in \mathbb{R}$.

Neste caso, a distribuição inerente ao intervalo $\hat{Y}(i)$ é também uma distribuição Uniforme.

Os parâmetros do *ID Model*, quando este assume uma distribuição Uniforme nos intervalos, podem ser determinados através da resolução de um problema de otimização quadrática baseado na distância de *Mallows*, com uma restrição de não negatividade nos parâmetros. Assim, o problema a minimizar poderá ser escrito da seguinte forma:

$$min \sum_{i=1}^{n} [\, (c_{Y(i)} - \sum_{j=1}^{p} (\alpha_j - \beta_j) c_{X_{ij}} - \gamma)^{\,2} + \frac{1}{3} (r_{Y(i)} - \sum_{j=1}^{p} (\alpha_j + \beta_j) r_{X_{ij}})^{\,2}] \quad (4.27)$$

com $-\alpha_j, -\beta_j \leq 0,\ j \in \{1, \dots, p\}$ e $\gamma \in \mathbb{R}$.

A solução proposta de utilizar uma função quantil de cada intervalo e do respetivo intervalo simétrico das p variáveis explicativas na previsão para a variável resposta, incrementa o número de parâmetros a estimar, sendo que, por cada variável explicativa, o modelo terá de estimar dois parâmetros. Assim, é importante que na utilização do modelo seja considerado um número de observações razoável e sempre superior ao número de parâmetros a estimar. Um problema similar também ocorre em alguns dos modelos apresentados no Capítulo 3 desta dissertação, em que para cada variável explicativa é necessário estimar um parâmetro para os centros e outro para os raios ou um parâmetro para o limite superior e outro para o limite inferior.

A partir dos parâmetros resultantes da minimização do problema (4.27), é possível definir o intervalo previsto. Considerando $t = 0$ obtemos o limite inferior e para $t = 1$ o limite superior do intervalo $\hat{Y}(i)$, resultando nas seguintes expressões:

$$\Psi^{-1}_{\hat{Y}(i)}(0) = \sum_{j=1}^{p} \alpha_j \left(c_{X_{ij}} - r_{X_{ij}}\right) - \sum_{j=1}^{p} \beta_j \left(c_{X_{ij}} + r_{X_{ij}}\right) + \gamma \quad (4.28)$$

$$\Psi^{-1}_{\hat{Y}(i)}(1) = \sum_{j=1}^{p} \alpha_j \left(c_{X_{ij}} + r_{X_{ij}}\right) - \sum_{j=1}^{p} \beta_j \left(c_{X_{ij}} - r_{X_{ij}}\right) + \gamma \quad (4.29)$$

Para cada observação i, o intervalo previsto é representado por:

$$I_{\hat{Y}(i)} = \left[\sum_{j=1}^{p} \left(\alpha_j l_{X_{ij}} - \beta_j u_{X_{ij}}\right) + \gamma, \sum_{j=1}^{p} \left(\alpha_j u_{X_{ij}} - \beta_j l_{X_{ij}}\right) + \gamma \right] \quad (4.30)$$

Através da equação (4.30) o centro e o raio dos intervalos previstos para a variável Y podem ser descritos através de uma relação linear com os centros e os raios, respetivamente, das variáveis explicativas X_j, com $j \in \{1, \dots, p\}$:

$$c_{\hat{Y}(i)} = \sum_{j=1}^{p} (\alpha_j - \beta_j) c_{X_{ij}} + \gamma; \qquad r_{\hat{Y}(i)} = \sum_{j=1}^{p} (\alpha_j + \beta_j) r_{X_{ij}} \tag{4.31}$$

com $\alpha_j, \beta_j \geq 0, j \in \{1, 2, \dots, p\}$ e $\gamma \in \mathbb{R}$.

Considerado a expressão que permite prever os centros e considerando que $\nu = \bar{Y} - \sum_{j=1}^{p} (\alpha_j - \beta_j) \bar{X}_j$, está provado que a soma dos desvios entre os centros dos intervalos observados e previstos é nula (Dias, 2014).

Observa-se através das expressões (4.31) que os parâmetros que definem as regressões lineares entre os centros e os raios são diferentes, mas estão relacionados. Apesar de com este modelo ser possível definir uma relação linear direta ou inversa, entre intervalos, a relação linear induzida entre os raios desses intervalos é sempre direta. A relação direta ou inversa entre as variáveis intervalares está relacionada com a relação linear entre os centros, isto é, uma variável intervalar X_j tem uma relação direta com a variável intervalar Y quando $\alpha_j > \beta_j$ e tem uma relação inversa se $\alpha_j < \beta_j$, com $j \in \{1, \dots, p\}$.

Exemplo:

De forma a ilustrar o modelo, será apresentado de seguida um pequeno exemplo. Recorrendo ao *R software* e ao programa desenvolvido no âmbito desta dissertação, serão analisados os dados da área queimada no parque natural de Montesinho. Este exemplo já foi analisado no âmbito do estudo de Dias (2014) e os dados poderão ser encontrados em Cortez e Morais (2007). Nesta base de dados está registada, por coordenada geográfica, a área queimada (hectares) desta região, de janeiro de 2000 a dezembro de 2003. É também conhecida a informação relativa a um conjunto de variáveis explicativas referentes à temperatura (medida em graus Celsius), o vento (medido em km/h) e a humidade (em percentagem). A variável resposta referente à área queimada será transformada em $\ln(Area + 1)$. A base de dados simbólica para este exemplo será

obtidas por agregação dos *microdados* por coordenada geográfica, e todas as variáveis aqui consideradas são intervalares.

Segundo o modelo apresentado nesta secção e através da resolução de um problema de otimização quadrático (MMQ) de forma a estimar os valores dos parâmetros, a relação linear que permite prever o $\ln(Area + 1)$ pelas variáveis intervalares referentes à temperatura, ao vento e à humidade é dada por:

$$\Psi^{-1}_{\widehat{Area}}(t) = 0.6406 + 0.0273\Psi^{-1}_{Temp}(t) + 0.3032\Psi^{-1}_{Vento}(t) - 0.0243\Psi^{-1}_{Vento}(1-t) + 0.0185\Psi^{-1}_{\%Humidade}(t) - 0.0133\Psi^{-1}_{\%Humidade}(1-t) \quad (4.32)$$

Os valores das medidas de ajuste deste modelo são as seguintes:

$\boldsymbol{\Omega}$	$\boldsymbol{RMSE_M}$	$\boldsymbol{RMSE_L}$	$\boldsymbol{RMSE_U}$
0.5041	0.9222	0.8103	1.1880

4.5 *ID Regression Model*: Distribuição Triangular Simétrica

Nesta secção será apresentado o *ID Model* quando uma Distribuição Triangular Simétrica é assumida nos intervalos. Esta extensão do modelo foi proposta em Dias e Brito (2017).

De acordo com as funções quantil definidas em (4.6) e (4.7), e assumindo uma distribuição Triangular Simétrica nos intervalos, é possível estabelecer uma relação linear entre a variável resposta *Y* e as variáveis explicativas X_j, com $j \in \{1, \dots, p\}$. Aplicando o *ID Model* a função quantil que representa o intervalo previsto (também com uma distribuição Triangular Simétrica) é dada por:

$$\Psi^{-1}_{\hat{Y}(i)}(t) = \begin{cases} \gamma + \sum_{j=1}^{p} \left[\left(\alpha_j - \beta_j\right) c_{X i_j} - \left(\alpha_j + \beta_j\right) r_{X_{ij}} \left(1 - \sqrt{2t}\right) \right], & 0 \le t \le \frac{1}{2} \\ \gamma + \sum_{j=1}^{p} \left[\left(\alpha_j - \beta_j\right) c_{X_{ij}} - \left(\alpha_j + \beta_j\right) r_{X_{ij}} \left(1 + \sqrt{2(1-t)}\right) \right], & \frac{1}{2} \le t \le 1 \end{cases} \quad (4.33)$$

Tal como no modelo em que é considerada uma distribuição Uniforme, os parâmetros para este caso são também obtidos através da minimização de um problema

de otimização, baseado na distância de *Mallows*, sujeito a restrições de não negatividade nos parâmetros:

$$min \sum_{i=1}^{n} [\,(c_{Y(i)} - \sum_{j=1}^{p} (\alpha_j - \beta_j) c_{X_{ij}} - \gamma)^{\,2} + \frac{1}{6} (r_{Y(i)} - \sum_{j=1}^{p} (\alpha_j + \beta_j) r_{X_{ij}})^{\,2}] \qquad (4.34)$$

com $-\alpha, -\beta_j \leq 0,\ i \in \{1, \dots, p\}$ e $\gamma \in \mathbb{R}$.

Utilizando os parâmetros obtidos através da resolução do problema (4.34), o cálculo do intervalo previsto, quando se assume uma distribuição Triangular Simétrica nos intervalos, é semelhante ao efetuado para o caso anterior. Substituindo $t = 0$ prevemos o limite inferior do intervalo e para $t = 1$ obtemos uma previsão para o limite superior do intervalo.

Este foi o primeiro desenvolvimento do modelo onde, assumindo que os valores possíveis dentro do intervalo não se distribuem uniformemente, foi possível captar uma maior variabilidade, principal objetivo na consideração de outras distribuições inerentes aos intervalos.

Exemplo

Recorrendo novamente aos dados apresentados na secção anterior e através da resolução de um problema de otimização quadrático (MMQ) de forma a obter os valores dos parâmetros, a relação linear que permite prever $\ln(Area + 1)$ pelas variáveis intervalares referentes à temperatura, ao vento e à humidade e considerando uma distribuição Triangular Simétrica é dada por:

$$\Psi^{-1}_{\widehat{Area}}(t) = 0.6385 + 0.0273\Psi^{-1}_{Temp}(t) + 0.3031\Psi^{-1}_{Vento}(t) - 0.0242\Psi^{-1}_{Vento}(1-t) + 0.0185\Psi^{-1}_{\%Humidade}(t) - 0.0134\Psi^{-1}_{\%Humidade}(1-t) \qquad (4.35)$$

Os valores das medidas de ajuste deste modelo são dados por:

$\boldsymbol{\Omega}$	$\boldsymbol{RMSE_M}$	$\boldsymbol{RMSE_L}$	$\boldsymbol{RMSE_U}$
0.3699	0.8969	0.8103	1.1880

4.6 ID Regression Model: Distribuição Triangular Geral

O principal objetivo proposto para esta dissertação é demonstrar a versatilidade do *ID Regression Model* adaptado ao contexto da análise simbólica, mais precisamente às variáveis simbólicas intervalares. Assim, nesta secção é apresentado o principal desenvolvimento proposto para o *ID Model*, onde será apresentado o modelo quando é assumida a distribuição Triangular Geral nos intervalos. Na Secção 4.5 já foi apresentado o modelo quando é assumida uma distribuição Triangular Simétrica, neste caso considera-se que a moda do intervalo coincide sempre com o seu centro

Nesta secção, o *ID Model* será adaptado a variáveis intervalares para as quais a distribuição inerente aos intervalos, que correspondem às suas observações, é uma distribuição Triangular Geral, em que a moda do intervalo poderá situar-se em qualquer valor entre o extremo inferior e o extremo superior do mesmo. O caso em que a moda assume um dos valores extremos do intervalo não será considerado nesta dissertação.

Ao ser considerado mais um ponto de referência em cada intervalo, a informação registada será mais rica e o *ID Model* ficará mais enriquecido. No entanto, ao acrescentar mais um ponto de referência na estrutura dos intervalos a complexidade da função quantil prevista vai aumentar, o que irá gerar dificuldades em prever os intervalos $\hat{Y}$. Por este motivo, nesta dissertação, será considerado apenas o caso da regressão linear simples, em que uma variável resposta é explicada por apenas uma variável explicativa.

Estando definida a função quantil de um intervalo (4.10) e a função quantil do intervalo simétrico (4.11) é possível representar a relação linear entre uma variável resposta e uma variável explicativa através da seguinte expressão:

$$\Psi^{-1}_{Y(i)}(t) = \alpha\Psi^{-1}_{X(i)}(t) - \beta\Psi^{-1}_{X(i)}(1-t) + \gamma \tag{4.36}$$

sendo γ o termo constante da regressão.

Dadas duas variáveis intervalares Y e X, onde $(u_{Y(i)}, l_{Y(i)}, m_{Y(i)})$ representam o limite superior, o limite inferior e a moda da variável resposta e $(u_{X(i)}, l_{X(i)}, m_{X(i)})$ que representam o limite superior, o limite inferior e a moda, respetivamente, da variável explicativa é possível definir uma relação linear, com base em (4.36), entre as variáveis, assumindo uma distribuição Triangular nos intervalos. Quando o ponto de mudança de

ramo das funções quantil $\Psi_X^{-1}(t)$ e $-\Psi_X^{-1}(1-t)$ não é o mesmo (o que só ocorre no caso particular da distribuição Triangular Simétrica), a combinação linear destas duas funções vai sempre dar origem a uma função definida por três ramos. Assim, e ao contrário dos casos anteriores, a combinação linear de intervalos com distribuição Triangular obtida através do *ID Model* não permite obter uma função que corresponda a uma função quantil da distribuição considerada.

Dependendo da localização do valor da moda no intervalo X_i, três situações podem ocorrer, tendo-se aplicado em cada uma delas a combinação linear definida em (4.36).

Situação 1:

Ocorre quando $\frac{m_{X(i)}-l_{X(i)}}{u_{X(i)}-l_{X(i)}} > \frac{u_{X(i)}-m_{X(i)}}{u_{X(i)}-l_{X(i)}}$. No caso desta condição se verificar, a função que resulta da combinação linear entre as funções $\Psi_X^{-1}(t)$ e $\Psi_X^{-1}(1-t)$, definidas com base em (4.10) e (4.11), é definida por:

$$\Phi_{\hat{Y}(i)}(t) = \begin{cases} \alpha l_{X(i)} - \beta u_{X(i)} + \left(\begin{matrix} \alpha\sqrt{\left(u_{X(i)} - l_{X(i)}\right)\left(m_{X(i)} - l_{X(i)}\right)} + \\ \beta\sqrt{\left(u_{X(i)} - l_{X(i)}\right)\left(u_{X(i)} - m_{X(i)}\right)} \end{matrix}\right)\sqrt{t} + \gamma, & 0 \le t \le \frac{u_{X(i)} - m_{X(i)}}{u_{X(i)} - l_{X(i)}} \\ (\alpha-\beta) l_{X(i)} + \sqrt{\left(u_{X(i)} - l_{X(i)}\right)\left(m_{X(i)} - l_{X(i)}\right)}\left(\alpha\sqrt{t} - \beta\sqrt{(1-t)}\right) + \gamma, & \frac{u_{X(i)} - m_{X(i)}}{u_{X(i)} - l_{X(i)}} < t < \frac{m_{X(i)} - l_{X(i)}}{u_{X(i)} - l_{X(i)}} \\ \alpha u_{X(i)} - \beta l_{X(i)} - \left(\begin{matrix} \alpha\sqrt{\left(u_{X(i)} - l_{X(i)}\right)\left(u_{X(i)} - m_{X(i)}\right)} + \\ \beta\sqrt{\left(u_{X(i)} - l_{X(i)}\right)\left(m_{X(i)} - l_{X(i)}\right)} \end{matrix}\right)\sqrt{1-t} + \gamma, & \frac{m_{X(i)} - l_{X(i)}}{u_{X(i)} - l_{X(i)}} \le t \le 1 \end{cases} \quad (4.37)$$

Esta função é contínua no intervalo $[0,1]$ e a sua derivada é positiva em todo o intervalo, e em consequência, é uma função não decrescente com um mínimo em $t = 0$ e um máximo em $t = 1$ e um único ponto de inflexão. No entanto, tal como já foi referido, esta função não representa um intervalo com uma distribuição Triangular.

Verificou-se que o valor da moda de um dado intervalo com distribuição Triangular é obtido no ponto de inflexão da função quantil que o representa. Assim, consideramos que o intervalo previsto para cada $Y(i)$ tem como extremo inferior o mínimo da função $\Phi_{\hat{Y}(i)}(t)$, obtido para $t = 0$, ou seja, $\Phi_{\hat{Y}(i)}(0) = \alpha l_{X(i)} - \beta u_{X(i)} + \gamma$, e como extremo superior o máximo da função $\Phi_{\hat{Y}(i)}(t)$, obtido para $t = 1$, $\Phi_{\hat{Y}(i)}(1) =$

$\alpha u_{X(i)} - \beta l_{X(i)} + \gamma$, e como moda $\Phi_{\hat{Y}(i)}(t^*)$, sendo t^* o valor onde a função $\Phi_{\hat{Y}(i)}(t)$ tem o seu único ponto de inflexão. Assim, nesta situação, a função quantil prevista para $Y(i)$ é dada por:

$$\Psi^{-1}_{\hat{Y}(i)}(t) = \begin{cases} l_{\hat{Y}(i)} + \sqrt{\left(u_{\hat{Y}(i)} - l_{\hat{Y}(i)}\right)\left(m_{\hat{Y}(i)} - l_{\hat{Y}(i)}\right)t}, & 0 \leq t \leq \dfrac{m_{\hat{Y}(i)} - l_{\hat{Y}(i)}}{u_{\hat{Y}(i)} - l_{\hat{Y}(i)}} \\ u_{\hat{Y}(i)} - \sqrt{\left(u_{\hat{Y}(i)} - l_{\hat{Y}(i)}\right)\left(u_{\hat{Y}(i)} - m_{\hat{Y}(i)}\right)(1-t)}, & \dfrac{m_{\hat{Y}(i)} - l_{\hat{Y}(i)}}{u_{\hat{Y}(i)} - l_{\hat{Y}(i)}} < t \leq 1 \end{cases} \quad (4.38)$$

com $l_{\hat{Y}(i)} = \alpha l_{X(i)} - \beta u_{X(i)} + \gamma$; $u_{\hat{Y}(i)} = \alpha u_{X(i)} - \beta l_{X(i)} + \gamma$ e $m_{\hat{Y}(i)} = \Phi_{\hat{Y}(i)}(t^*)$.

Neste primeiro caso, dependendo dos valores dos parâmetros, temos três valores possíveis para a moda.

Se $\frac{\beta\left(u_{X(i)} - m_{X(i)}\right)^{3/2}}{\left(m_{X(i)} - l_{X(i)}\right)^{3/2}} < \alpha < \frac{\beta\left(m_{X(i)} - l_{X(i)}\right)^{3/2}}{\left(u_{X(i)} - m_{X(i)}\right)^{3/2}}$, o ponto de inflexão de $\Phi_{\hat{Y}(i)}(t)$ será obtido em $t^* = \frac{\alpha^{2/3}}{\alpha^{2/3} + \beta^{2/3}}$. A moda do intervalo previsto neste caso é $\Phi_{\hat{Y}(i)}(t^*) = (\alpha - \beta) l_{X(i)} + \sqrt{\left(u_{X(i)} - l_{X(i)}\right)\left(m_{X(i)} - l_{X(i)}\right)}\left(\frac{\alpha^{4/3} - \beta^{4/3}}{\sqrt{\alpha^{2/3} + \beta^{2/3}}}\right) + \gamma$.

Se $\alpha < \frac{\beta\left(u_{X(i)} - m_{X(i)}\right)^{3/2}}{\left(m_{X(i)} - l_{X(i)}\right)^{3/2}}$ o ponto de inflexão será obtido em $t^* = \frac{u_{X(i)} - m_{X(i)}}{u_{X(i)} - l_{X(i)}}$, sendo que substituindo t^* em (4.37) é obtida a moda do intervalo previsto, ou seja, $\Phi_{\hat{Y}(i)}(t^*) = \alpha l_{X(i)} - \beta m_{X(i)} + \alpha\sqrt{\left(u_{X(i)} - m_{X(i)}\right)\left(m_{X(i)} - l_{X(i)}\right)} + \gamma$.

Se $\alpha > \frac{\beta\left(m_{X(i)} - l_{X(i)}\right)^{3/2}}{\left(u_{X(i)} - m_{X(i)}\right)^{3/2}}$ o ponto de inflexão será obtido em $t^* = \frac{m_{X(i)} - l_{X(i)}}{u_{X(i)} - l_{X(i)}}$, sendo a moda do intervalo previsto dada por: $\Phi_{\hat{Y}(i)}(t^*) = \alpha m_{X(i)} - \beta l_{X(i)} + \beta\sqrt{\left(u_{X(i)} - m_{X(i)}\right)\left(m_{X(i)} - l_{X(i)}\right)} + \gamma$.

Situação 2:

Ocorre se $\frac{m_{X(i)} - l_{X(i)}}{u_{X(i)} - l_{X(i)}} < \frac{u_{X(i)} - m_{X(i)}}{u_{X(i)} - l_{X(i)}}$. No caso desta condição se verificar a função que resulta da relação linear entre a variável resposta e explicativa, com base em (4.10) e (4.11), é definida por:

$$\Phi_{\hat{Y}(i)}(t)=\begin{cases}\alpha l_{X(i)}-\beta u_{X(i)}+\left(\begin{matrix}\alpha\sqrt{\left(u_{X(i)}-l_{X(i)}\right)\left(m_{X(i)}-l_{X(i)}\right)}+\\ \beta\sqrt{\left(u_{X(i)}-l_{X(i)}\right)\left(u_{X(i)}-m_{X(i)}\right)}\end{matrix}\right)\sqrt{t}+\gamma, & 0\le t\le\frac{u_{X(i)}-m_{X(i)}}{u_{X(i)}-l_{X(i)}}\\ (\alpha-\beta)u_{X(i)}+\sqrt{\left(u_{X(i)}-l_{X(i)}\right)\left(u_{X(i)}-m_{X(i)}\right)}\left(\beta\sqrt{t}-\alpha\sqrt{(1-t)}\right)+\gamma, & \frac{u_{X(i)}-m_{X(i)}}{u_{X(i)}-l_{X(i)}}<t<\frac{m_{X(i)}-l_{X(i)}}{u_{X(i)}-l_{X(i)}}\\ \alpha u_{X(i)}-\beta l_{X(i)}-\left(\begin{matrix}\alpha\sqrt{\left(u_{X(i)}-l_{X(i)}\right)\left(u_{X(i)}-m_{X(i)}\right)}+\\ \beta\sqrt{\left(u_{X(i)}-l_{X(i)}\right)\left(m_{X(i)}-l_{X(i)}\right)}\end{matrix}\right)\sqrt{1-t}+\gamma, & \frac{m_{X(i)}-l_{X(i)}}{u_{X(i)}-l_{X(i)}}\le t\le 1\end{cases}\quad(4.39)$$

Tal como na situação anterior, definimos a função quantil prevista para $Y(i)$ a partir dos extremos de $\Phi_{\hat{Y}(i)}(t)$ e do seu ponto de inflexão, que nos permitirá determinar a moda do intervalo, ou seja:

$$\Psi^{-1}_{\hat{Y}(i)}(t)=\begin{cases}l_{\hat{Y}(i)}+\sqrt{\left(u_{\hat{Y}(i)}-l_{\hat{Y}(i)}\right)\left(m_{\hat{Y}(i)}-l_{\hat{Y}(i)}\right)t}, & 0\le t\le\dfrac{m_{\hat{Y}(i)}-l_{\hat{Y}(i)}}{u_{\hat{Y}(i)}-l_{\hat{Y}(i)}}\\ u_{\hat{Y}(i)}-\sqrt{\left(u_{\hat{Y}(i)}-l_{\hat{Y}(i)}\right)\left(u_{\hat{Y}(i)}-m_{\hat{Y}(i)}\right)(1-t)}, & \dfrac{m_{\hat{Y}(i)}-l_{\hat{Y}(i)}}{u_{\hat{Y}(i)}-l_{\hat{Y}(i)}}<t\le 1\end{cases}\quad(4.40)$$

com $l_{\hat{Y}(i)}=\alpha l_{X(i)}-\beta u_{X(i)}+\gamma$; $u_{\hat{Y}(i)}=\alpha u_{X(i)}-\beta l_{X(i)}+\gamma$ e $m_{\hat{Y}(i)}=\Phi_{\hat{Y}(i)}(t^*)$.

Também nesta situação, temos três valores possíveis para a moda.

Se $\frac{\beta\left(m_{X(i)}-l_{X(i)}\right)^{3/2}}{\left(u_{X(i)}-m_{X(i)}\right)^{3/2}}<\alpha<\frac{\beta\left(u_{X(i)}-m_{X(i)}\right)^{3/2}}{\left(m_{X(i)}-l_{X(i)}\right)^{3/2}}$, o ponto de inflexão de $\Phi_{\hat{Y}(i)}(t)$ será obtido em $t^*=\frac{\beta^{2/3}}{\alpha^{2/3}+\beta^{2/3}}$. A moda do intervalo previsto neste caso é $\Phi_{\hat{Y}(i)}(t^*)=(\alpha-\beta)u_{X(i)}+\sqrt{\left(u_{X(i)}-l_{X(i)}\right)\left(u_{X(i)}-m_{X(i)}\right)}\left(\frac{\beta^{4/3}-\alpha^{4/3}}{\sqrt{\alpha^{2/3}+\beta^{2/3}}}\right)+\gamma$.

Se $\alpha<\frac{\beta\left(m_{X(i)}-l_{X(i)}\right)^{3/2}}{\left(u_{X(i)}-m_{X(i)}\right)^{3/2}}$ o ponto de inflexão será obtido em $t^*=\frac{u_{X(i)}-m_{X(i)}}{u_{X(i)}-l_{X(i)}}$, sendo que substituindo t^* em (4.39) é obtida a moda do intervalo previsto, ou seja, $\Phi_{\hat{Y}(i)}(t^*)=\alpha u_{X(i)}-\beta m_{X(i)}+\alpha\sqrt{\left(u_{X(i)}-m_{X(i)}\right)\left(m_{X(i)}-l_{X(i)}\right)}+\gamma$.

Se $\alpha>\frac{\beta\left(u_{X(i)}-m_{X(i)}\right)^{3/2}}{\left(u_{X(i)}-l_{X(i)}\right)^{3/2}}$ o ponto de inflexão será obtido em $t^*=\frac{m_{X(i)}-l_{X(i)}}{u_{X(i)}-l_{X(i)}}$, sendo a moda do intervalo previsto dada por: $\Phi_{\hat{Y}(i)}(t^*)=\alpha m_{X(i)}-\beta u_{X(i)}+\beta\sqrt{\left(u_{X(i)}-m_{X(i)}\right)\left(m_{X(i)}-l_{X(i)}\right)}+\gamma$.

Situação 3:

Ocorre quando $\frac{m_{X(i)}-l_{X(i)}}{u_{X(i)}-l_{X(i)}} = \frac{u_{X(i)}-m_{X(i)}}{u_{X(i)}-l_{X(i)}}$. Neste cenário a moda é igual ao centro do intervalo, tornando-se apenas necessário estabelecer uma relação linear para o limite inferior e para o limite superior. É possível assumir esta hipótese porque este modelo está definido para apenas uma variável explicativa. Estamos no caso da distribuição Triangular Simétrica.

Não sendo possível utilizar uma otimização quadrática para obter os parâmetros como nos modelos anteriores, estes terão de ser obtidos através da resolução de um problema de otimização numérica. Isto é, os resultados que irão ser obtidos para os parâmetros do modelo não serão os parâmetros ótimos, mas serão os melhores valores para os parâmetros.

O algoritmo que irá ser utilizado na otimização numérica foi desenvolvido por Cheira et al (2017). Este irá permitir a obtenção dos melhores parâmetros para a relação linear do modelo através da minimização da soma dos quadrados da distância de *Mallows* definida por Cheira et al (2017), entre os intervalos observados e os intervalos previstos. Este algoritmo irá procurar iterativamente, valores para os parâmetros até que o valor da função a minimizar estabilize num valor mínimo.

Assim, para a obtenção dos parâmetros do modelo será necessário escolher valores iniciais para os parâmetros $\alpha, \beta \, e \, \gamma$ e definir a função, dependente desses parâmetros, que se pretende minimizar. A função a minimizar será a soma do quadrado da distância de *Mallows* entre os intervalos observados e os intervalos previstos para cada indivíduo i, $Y(i)$ e $\hat{Y}(i)$. Assim, a função a otimizar será a seguinte:

$$min \sum_{i=1}^{n} D_M^2 \left(\Psi_Y^{-1}(i), \Psi_{\hat{Y}}^{-1}(i) \right) \tag{4.41}$$

Utilizando os parâmetros obtidos através da resolução do problema de minimização numérica é possível definir os intervalos previstos.

Com este modelo foi realizado um desenvolvimento importante no contexto da

Análise Simbólica de Dados e mais concretamente nos modelos de regressão linear para variáveis intervalares. Ao considerar a distribuição Triangular nos intervalos é possível captar mais informação, aproximar os intervalos da realidade dos dados e enriquecer as relações lineares entre intervalos.

Exemplo

Este modelo foi desenvolvido para uma relação linear entre duas variáveis, sendo que não é possível utilizar o mesmo número de variáveis que no exemplo das duas secções anteriores. Assim, recorrendo apenas a parte dos dados apresentados na secção anterior e através da resolução de um problema de otimização númerico de forma a obter os valores dos melhores parâmetros, a relação linear entre $\ln(Area + 1)$ e a variável explicativa vento é dada por:

$$\Psi^{-1}_{\widehat{Area}}(t) = 1.4790 + 0.4428\Psi^{-1}_{Vento}(t) - 0.2165\Psi^{-1}_{Vento}(1-t) \qquad (4.42)$$

Os valores das medidas de ajuste deste modelo são dadas por:

$\boldsymbol{\Omega}$	$\boldsymbol{RMSE_M}$	$\boldsymbol{RMSE_L}$	$\boldsymbol{RMSE_U}$
-	0.9193	0.8965	1.2130

Capítulo 5

Implementação e Aplicações do *ID Model*

Neste capítulo será apresentada a implementação em *R software* do *ID Regression Model*, como proposto no planeamento desta dissertação, onde será detalhada a forma de *inputs* dos dados, a estrutura do código e a leitura dos respetivos *outputs*. Posteriormente, será utilizada uma base de dados real, referente ao consumo de eletricidade de um estabelecimento comercial, de forma a ilustrar a aplicação do *ID Model*, com diferentes distribuições, e comparar o seu desempenho face aos restantes modelos apresentados no Capítulo 3. O desempenho dos modelos será avaliado de acordo com as medidas de qualidade apresentadas no Capítulo 4.

5.1 Implementação do Modelo

Nesta secção será detalhada a implementação do modelo em estudo, o formato de *input* dos dados para que os mesmos sejam reconhecidos, as funções desenvolvidas no *software R* e a obtenção e leitura dos *outputs.*

Como é necessário transformar os dados clássicos em dados simbólicos, agregando as observações a níveis superiores ("grupos"), foi criada uma função, denominada por "*ToSimbolic(x, m)*", em *R software* para realizar este processo. Nesta função pode ser apenas considerada uma variável. Para que a função realize a transformação são necessários dois parâmetros: o parâmetro *x* que é referente aos dados clássicos e o parâmetro *m* referente à distribuição a considerar. Os dados devem ser carregados em duas colunas: a primeira com a indicação do "grupo" a que pertence a observação e a segunda com os valores referentes a cada observação, como ilustrado na Figura 5.1, em que é desejado que a informação registada seja agregada ao dia. Se o parâmetro *m* for substituído por "1" ou "2", o *output* será uma matriz de duas colunas: a primeira com os centros e a segunda com os raios dos intervalos para cada um dos "grupos". Se for substituído por "3", será acrescentada na matriz do *output* final mais uma coluna com a moda dos intervalos. Para conjuntos de dados com mais do que uma

variável, foi desenvolvida a função *"ToSimbolic2(x,m)"* que aplica a função *"ToSimbolic(x,m)"* a cada coluna de dados introduzidos, sendo que o *output* será uma matriz com os pontos de referência de cada intervalo, para cada uma das variáveis.

Nível a Agregar	*Valor Registado*
06/07/2016	7,9
06/07/2016	8,7
06/07/2016	7,4
07/07/2016	7,7
07/07/2016	7,6
07/07/2016	7,7
07/07/2016	7,5
08/07/2016	7
08/07/2016	7,7
08/07/2016	7,3
09/07/2016	11,1
09/07/2016	11,1
09/07/2016	9,3
(...)	(...)

Figura 5.1: Exemplo de uma tabela de dados clássicos, para transformação em intervalos pela função *"ToSimbolic"* em *R software*

O objetivo do *package* desenvolvido é a aplicação do *ID Model* a variáveis intervalares, de acordo com uma distribuição nos dados intervalares. Para tal é necessário obter os parâmetros para os modelos de regressão linear e as previsões para os intervalos da variável resposta. Assim, de forma a que o programa desenvolvido reconheça os dados utilizados, os mesmos necessitam de um pré-processamento, caso estes não sejam carregados pela função anterior.

O ficheiro com os dados deve ser carregado em *R software* de forma a que as tabelas assumam o formato de *data.frame* ou de uma matriz, sendo que com a aplicação da primeira função do modelo programado, todos os dados serão convertidos em matrizes. Os dados referentes à variável resposta e os dados referentes às variáveis explicativas devem ser carregados em ficheiros separados e utilizados no modelo em tabelas simbólicas separadas. O ficheiro referente aos dados da variável resposta apenas necessitará de ter duas colunas, a primeira referente os centros dos intervalos e a segunda referente aos raios, sendo que caso seja utilizado o modelo quando este assume uma distribuição Triangular Geral é necessário acrescentar uma terceira coluna com a moda de cada intervalo, sendo que a mesma é calculada através da equação da moda de King.

O ficheiro referente aos dados das variáveis explicativas deverá conter duas colunas por cada variável, a primeira referente ao centro do intervalo e a segunda referente ao raio, sendo necessário acrescentar uma terceira coluna com a informação da moda, caso os intervalos assumam uma distribuição Triangular Geral. Assim, os valores (intervalares) das variáveis serão representados nas colunas pela seguinte ordem: $c_1, r_1, m_1^*, \dots, c_p, r_p, m_p^*$ (*só é necessário incluir a moda, m_p, no caso de se assumir uma distribuição Triangular Geral nos intervalos observados). Na Figura 5.2 é apresentado um exemplo de como os dados simbólicos, no formato Excel, deverão ser carregados no *R software*, no caso de a distribuição inerente aos intervalos ser uma distribuição Triangular Geral.

	A	B	C	D	E	F
1	**Centro**	**Raio**	**Moda**	**Centro**	**Raio**	**Moda**
2	27,45	5,65	31	34,2	5,1	37,6
3	28	8,9	22,2	32,75	7,45	30,3
4	25,4	9	34	31,5	8,6	27,1
5	27,35	9,35	33,6	32,2	8,1	39,1
6	25,35	7,25	31,8	31,95	7,55	25,9
7	20,2	7,6	25,2	28,2	7,3	25,2
8	20,05	8,45	12,9	28,2	7,7	22,3
9	21,45	10,45	13,3	29,7	9,1	21,4
10	24,15	9,35	16,7	30,25	9,75	23,5
11	25,65	8,25	17,7	31,8	8,3	28,1
12	27,3	10,4	18,4	31,75	8,55	26,9
13	30,6	10	22	33,65	6,95	39,9
14	30,6	7,7	24,5	34,8	5,5	38,9
15	27,8	9,7	20,8	32,95	7,65	39,3
16	24,5	9,5	16,7	30,65	8,85	25,2

Figura 5.2: Exemplo de uma matriz "*input*" de dados simbólicos

O *package* foi implementado em cinco funções principais: uma função compiladora de todo o processo, que contempla um conjunto de funções para cada uma das distribuições estudadas; uma função para o cálculo dos parâmetros obtidos para o modelo de regressão linear; uma função para realizar as previsões dos intervalos com base nos parâmetros do modelo e uma função que contempla as várias medidas que permitem avaliar a qualidade do modelo. Mais ainda, foram implementadas funções auxiliares que permitem o cálculo das distâncias de *Mallows* para variáveis com uma distribuição Uniforme, Triangular Simétrica e Triangular Geral; uma função de otimização numérica, sendo que a otimização quadrática foi efetuada recorrendo ao *package "quadprog"*, já incluído nos *packages* disponíveis no *R software.*

O *package* desenvolvido requer um conjunto de parâmetros iniciais: uma matriz que contém a informação da variável resposta, uma matriz que contém a informação das

variáveis explicativas e a indicação da distribuição inerente aos intervalos. Para esta última, "1" corresponde ao modelo com uma distribuição Uniforme, "2" ao modelo com uma distribuição Triangular Simétrica e "3" ao modelo com uma distribuição Triangular Geral. Nos casos da distribuição Uniforme e Triangular Simétrica, toda a programação foi realizada com base na utilização de matrizes, ou seja, os *inputs Y, X* e toda a informação será processada no formato matricial.

Para o cálculo dos parâmetros do modelo de regressão linear, se o modelo a utilizar for "1" ou "2" é efetuada uma otimização quadrática com restrições de não negatividade (Dias, 2014), recorrendo ao *package "quadprog".* No caso em que é assumida uma distribuição Triangular Geral, "3", a obtenção dos parâmetros será realizada através de uma otimização numérica. Para este último caso é necessário introduzir parâmetros iniciais na função, por defeito são considerados $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 0.5$, e considerar a restrição de não negatividade para α e β. No entanto os parâmetros poderão ser alterados acrescentando um vetor $p = c(\alpha, \beta, \gamma)$ com os valores desejados, no quarto parâmetro da função inicial ("*Modelo(m, Y, X, p)"*). Os primeiros intervalos previstos por esta função serão obtidos a partir dos parâmetros inicialmente introduzidos. Dado que neste caso estamos a aplicar uma otimização numérica, será efetuada uma "procura sistemática" de valores para os parâmetros do modelo, até serem encontrados os valores que minimizam a soma dos quadrados da distância de *Mallows* entre os intervalos previstos e observados da variável resposta.

As previsões para os intervalos da variável resposta são obtidas através de uma função que faz parte do *package* e que, a partir dos parâmetros ótimos, determina esses intervalos. Foi também programada uma função que, a partir das previsões obtidas, determina os valores das medidas de qualidade do modelo: Ω, $RMSE_L$, $RMSE_U$ e $RMSE_M$. O *output* final devolve ao utilizador os valores dos parâmetros, os intervalos previstos para a variável resposta e as medidas que avaliam a qualidade do modelo aplicado.

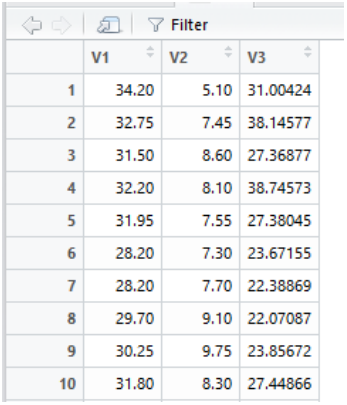
O código *R* desenvolvido no âmbito desta dissertação poderá ser consultado com um maior detalhe no Anexo desta dissertação.

5.2 Aplicação do modelo a um caso real

Nesta secção são apresentados os resultados obtidos pela aplicação do *ID Model* a uma base de dados real, quando se assume uma distribuição Triangular Geral nos intervalos. Posteriormente, será realizada uma comparação com os resultados obtidos por métodos já existentes e apresentados na revisão bibliográfica, métodos estes que estão incluídos no *package iRegression* desenvolvido por Lima Neto (Lima Neto e De Carvalho 2008). Será realizada ainda uma comparação dos resultados obtidos pela aplicação do *ID Model* quando nos intervalos é assumida uma distribuição Uniforme ou uma distribuição Triangular Simétrica.

Os dados que irão ser utilizados neste estudo referem-se ao consumo de eletricidade, em *kWh*, de uma central de frio positivo de uma loja de retalho alimentar. Os dados do consumo de eletricidade são referentes a uma loja situada no interior de Portugal Continental, local onde se registam: temperaturas exteriores à loja elevadas e de grande amplitude térmica nos períodos de Verão; e temperaturas exteriores negativas e de menor amplitude térmica nos períodos de Inverno. A base de dados é composta por 34.842 observações, referentes a uma sequência de períodos de 15 minutos durante 366 dias (6 de julho de 2016 a 6 de julho de 2017). Para cada um destes períodos é registado o consumo da central de frio positivo da loja e a temperatura exterior. O consumo da central de frio positivo refere-se ao consumo dos expositores, de frio positivo, da área de venda e às câmaras frigoríficas da loja. O consumo de arcas congeladoras e outros aparelhos de frio negativo não estão refletidos neste consumo em estudo.

O objetivo é estudar a relação linear dos consumos da loja com base na temperatura exterior à mesma, medida em graus Celsius. De forma a considerar a variabilidade nos dados de consumo e da temperatura, os mesmos foram agregados ao dia. A base de dados simbólica é composta por 366 observações, que representam cada um dos dias do ano, onde cada dia agrega em cada variável todas as observações de 15 minutos da base de dados inicial. A agregação dos dados foi realizada através da função "*ToSimbolic2*" (apresentada na secção 5.1). Na Figura 5.3, estão apresentadas as primeiras observações agregadas. Para a agregação dos dados contínuos em intervalos foi utilizada a moda de King para calcular a moda de cada intervalo.



	V1	V2	V3
1	34.20	5.10	31.00424
2	32.75	7.45	38.14577
3	31.50	8.60	27.36877
4	32.20	8.10	38.74573
5	31.95	7.55	27.38045
6	28.20	7.30	23.67155
7	28.20	7.70	22.38869
8	29.70	9.10	22.07087
9	30.25	9.75	23.85672
10	31.80	8.30	27.44866

Figura 5.3: Variável Resposta Consumo (kWh) no formato de uma tabela simbólica

Tal como já foi mencionado, existe uma diferença da amplitude térmica entre o período do Inverno e o período do Verão, motivo pelo qual serão realizados dois estudos para avaliar o desempenho do modelo. No primeiro estudo será aplicado apenas um modelo a todos os dias do ano de forma a obter os parâmetros que permitam prever o intervalo de consumos. No segundo estudo serão aplicados dois modelos, um ao período de maior calor e maior amplitude térmica e o segundo ao período mais frio e de menor amplitude térmica.

Estudo 1

Como mencionado, neste primeiro estudo será aplicado o modelo à base de dados simbólica considerando os registos anuais. Assim, aplicando o *ID Model*, e assumindo uma distribuição Triangular Geral nos intervalos, a relação linear entre a variável resposta *Cons* e a variável explicativa *Temp* é dada por:

$$\Psi^{-1}_{\widehat{Cons}}(t) = 3.9524 + 0.3153\Psi^{-1}_{Temp}(t) - 0.0680\Psi^{-1}_{Temp}(1-t) \qquad (5.1)$$

Pela análise dos parâmetros obtidos para o modelo, é possível concluir que a relação que existe entre o consumo de eletricidade da central de frio positivo e a temperatura exterior à loja é direta, porque o valor do parâmetro α é superior ao valor do parâmetro β.

Na Tabela 5.1 são apresentados os valores das medidas de avaliação deste modelo:

Tabela 5.1: Medidas de avaliação do modelo

Modelo	$\boldsymbol{RMSE_M}$	$\boldsymbol{RMSE_L}$	$\boldsymbol{RMSE_U}$
$ID_{Triangular}$	0.6330	0.8099	0.9605

Abaixo, e de forma a ser possível comparar os resultados do modelo desenvolvido no âmbito desta dissertação, são colocados lado a lado, os intervalos observados e os intervalos previstos através do *ID Model*, quando a distribuição Triangular é considerada. Esta análise aos resultados é realizada para 2 meses do ano, um mês no verão, onde a amplitude das temperaturas tem uma tendência a ser maior, e um mês do período de Inverno, onde a amplitude dos intervalos da temperatura tende a ser menor.

Na Figura 5.4 são apresentados os intervalos observados e os intervalos previstos, para cada dia, do mês de agosto. Em cada par de barras do gráfico, a primeira é referente ao intervalo observado para um dado dia, com o respetivo limite superior, limite inferior e a moda do intervalo representada por um traço de cor verde e a segunda barra representa o respetivo intervalo previsto, para o mesmo dia, sendo que a moda é representada por um traço a vermelho. Os intervalos estimados apresentam um ajuste razoável no limite inferior e superior dos intervalos, no entanto existiu alguma dificuldade em prever a moda para alguns dias deste mês.

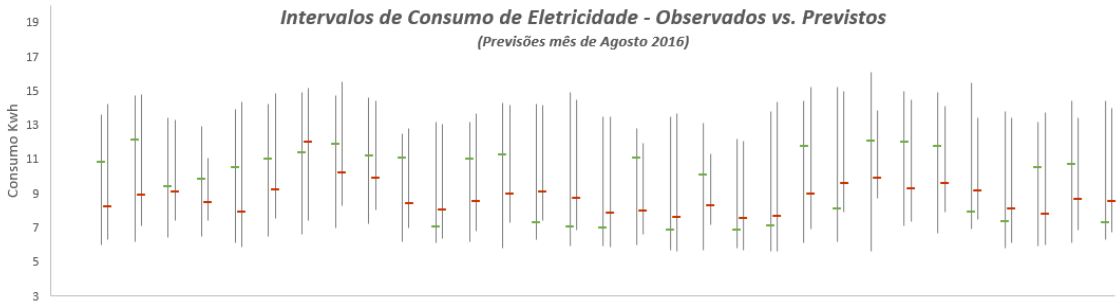


Figura 5.4: Representação dos intervalos observados e previstos no mês de agosto

Na Figura 5.5 é apresentada a mesma análise para um mês em que as temperaturas exteriores são mais baixas e a amplitude entre a temperatura máxima e mínima é menor.

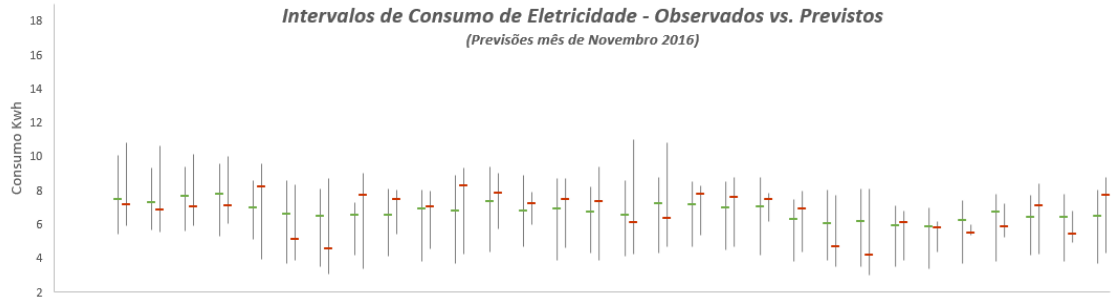


Figura 5.5: Representação dos intervalos observados e previstos no mês de novembro

No mês de novembro, o ajuste entre os intervalos observados e previstos mantém um bom desempenho nos limites superiores e inferiores, salvaguardando novamente algumas observações. No entanto, pela observação da Figura 5.4 e 5.5, a diferença entre a moda observada e a moda estimada, é menor no mês de novembro face ao mês de agosto.

De forma a comparar os resultados obtidos, este conjunto de dados foi também aplicado ao *ID Model*, quando se assumem outras distribuições e aos modelos apresentados no Capítulo 3: o Método do Centro (CM) (Billard e Diday 2000); o Método do Mínimo e Máximo (MinMax) (Billard e Diday 2002); o Método do Centro e Raio (CRM) (Lima Neto e De Carvalho 2008); o Método do Centro e Raio com Restrição (CCRM) (Lima Neto e De Carvalho 2010); o *ID Model* considerando distribuição Uniforme nos intervalos ($ID_{Uniforme}$) (Dias e Brito 2017); o *ID Model* considerando distribuição Triangular Simétrica nos intervalos ($ID_{TriangularSimetrica}$) (Dias e Brito 2017). A relação linear resultante de cada um dos modelos e os valores das respetivas medidas de qualidade estão apresentadas na Tabela 5.2 e na Tabela 5.3, respetivamente.

Tabela 5.2: Relações lineares para cada um dos modelos: aplicação anual

Modelos	***Expressões que permitem prever os intervalos***
$ID_{Uniforme}$	$\Psi^{-1}_{\widehat{Cons}}(t) = 3.7490 + 0.3106\Psi^{-1}_{Temp}(t) - 0.0709\Psi^{-1}_{Temp}$(1-t)
$ID_{TriangularSimetrica}$	$\Psi^{-1}_{\widehat{Cons}}(t) = 3.7490 + 0.3106\Psi^{-1}_{Temp}(t) - 0.0709\Psi^{-1}_{Temp}$(1-t)
$ID_{Triangular}$	$\Psi^{-1}_{\widehat{Cons}}(t) = 3.9524 + 0.3153\Psi^{-1}_{Temp}(t) - 0.0680\Psi^{-1}_{Temp}$(1-t)
CM	$c_{\widehat{Cons}} = 3.7490 + 0.2397c_{Temp}$
$MinMax$	$u_{\widehat{Cons}} = 4.4138 + 0.1834u_{Temp}$ e $l_{\widehat{Cons}} = 3.1652 + 0.1834l_{Temp}$
CRM	$c_{\widehat{Cons}} = 3.7490 + 0.2397c_{Temp}$ e $r_{\widehat{Cons}} = 1.2527 + 0.2255c_{Temp}$
$CCRM$	$c_{\widehat{Cons}} = 3.7490 + 0.2397c_{Temp}$ e $r_{\widehat{Cons}} = 1.2527 + 0.2255c_{Temp}$

Através dos parâmetros obtidos anteriormente, foram calculadas as medidas de qualidade para cada um dos modelos, que são apresentadas na Tabela 5.3.

Tabela 5.3: Medidas de Qualidade para cada um dos modelos: aplicação anual

Modelos	$\boldsymbol{\Omega}$	$\boldsymbol{RMSE_M}$	$\boldsymbol{RMSE_L}$	$\boldsymbol{RMSE_U}$
$ID_{Uniforme}$	0.9416	0.6020	0.6789	0.9431
$ID_{TriangularSimetrica}$	0.9399	0.5331	0.6789	0.9431
$ID_{Triangular}$	-	0.6330	0.8099	0.9605
CM	0.7611	0.6908	1.2831	1.4209
$MinMax$	0.8857	0.5080	0.4657	0.8126
CRM	0.7198	0.7825	0.4351	0.6107
$CCRM$	0.7198	0.7825	0.4351	0.6107

Os modelos obtidos são idênticos para o *ID Model* quando este assume uma distribuição Uniforme e uma distribuição Triangular Simétrica. Isto significa que a consideração de uma distribuição com moda no centro do intervalo não levou a quaisquer alterações na relação linear entre as variáveis.

Os parâmetros dos modelos do Centro e Raio e do Centro e Raio com Restrições são iguais, o que indica que existe uma relação direta entre os raios das variáveis e que por isso a restrição imposta no modelo não tem qualquer influência na construção do mesmo.

Analisando as medidas de qualidade dos modelos, o valor de Ω é de aproximadamente 94% no *ID Model* quando é assumida uma distribuição Uniforme ou uma distribuição Triangular Simétrica, sendo que no segundo caso é ligeiramente inferior.

Este valor significa que 94% da variação total do consumo de eletricidade é explicada pela temperatura exterior à loja. Assim, conhecendo as previsões das temperaturas para o exterior da loja, será possível prever qual o consumo mínimo e o consumo máximo da loja, para um dia.

Segundo as medidas de qualidade Ω e $RMSE_M$, os modelos com um pior ajuste aos consumos observados neste estudo são o modelo do Centro e do Raio e o modelo do Centro e do Raio com Restrições. No entanto, pelas medidas $RMSE_L$ e $RMSE_U$, o modelo que apresentou um pior desempenho foi o modelo do Centro.

Estudo 2

Neste segundo estudo o problema anterior será dividido em dois sub-problemas. Será aplicado o modelo às observações referentes ao período de mais frio, de 1 de outubro de 2016 a 31 de março de 2017, e posteriormente às observações referentes ao período de maior calor, de 6 de julho de 2016 a 31 de setembro de 2016 e de 1 de abril de 2017 a 6 de julho de 2017.

Aplicando o *ID Model*, assumindo uma distribuição Triangular Geral nos intervalos, a relação linear entre a variável resposta *Cons* e a variável explicativa *Temp* para cada uma das situações é dada por:

$$\Psi^{-1}_{\widehat{Cons_{Quente}}}(t) = 3.9218 + 0.3232\Psi^{-1}_{Temp_{Quente}}(t) - 0.0699\Psi^{-1}_{Temp_{Quente}}(1-t) \quad (5.2)$$

$$\Psi^{-1}_{\widehat{Cons_{Frio}}}(t) = 4.5607 + 0.2689\Psi^{-1}_{Temp_{Frio}}(t) - 0.0845\Psi^{-1}_{Temp_{Frio}}(1-t) \quad (5.3)$$

Pela análise dos parâmetros obtidos para os modelos, também é possível concluir que a relação que existe entre o consumo de eletricidade das centrais de frio positivo e a temperatura exterior à loja é direta, porque o valor do parâmetro α é superior ao valor do parâmetro β em ambos os modelos.

Na Tabela 5.4 são apresentados os valores das medidas de avaliação para os dois modelos:

Tabela 5.4: Medidas de avaliação do modelo

Modelo	$RMSE_M$	$RMSE_L$	$RMSE_U$
$ID_{Triangular\_Quente}$	0.6754	0.7269	0.9624
$ID_{Triangular\_Frio}$	0.5002	0.9051	0.8778

Abaixo, de forma a ser possível comparar os resultados do *ID Model* para a distribuição Triangular Geral, são colocados lado a lado, os intervalos observados e os intervalos previstos. Esta análise aos resultados é realizada para os dois modelos estudados e para uma sequência de dias aleatória dentro de cada um dos períodos.

Assim, na Figura 5.6 são apresentados os intervalos observados e previstos, para uma sequência de dias do período quente. Em cada par de barras do gráfico, a primeira é referente ao intervalo observado para uma observação, com o respetivo limite superior, limite inferior e a moda do intervalo representada por um traço de cor verde e a segunda barra refere-se ao intervalo previsto para a mesma observação, sendo que a moda é agora representada por um traço a vermelho. Com base no gráfico da Figura 5.6, verifica-se que os intervalos previstos apresentam um bom ajuste no limite inferior e superior dos intervalos.

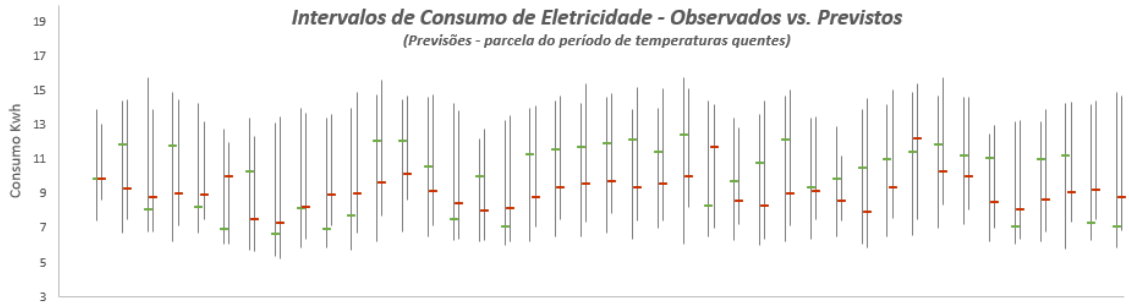


Figura 5.6: Representação dos intervalos observados e previstos para um período quente

Na Figura 5.7 é apresentada a mesma análise, mas para uma sequência de dias do período frio. Nestes resultados existem dificuldades do modelo em realizar previsões corretas quando o intervalo observado do consumo é de baixa amplitude. No entanto, existem observações, com uma boa previsão para os extremos e para a moda dos intervalos.

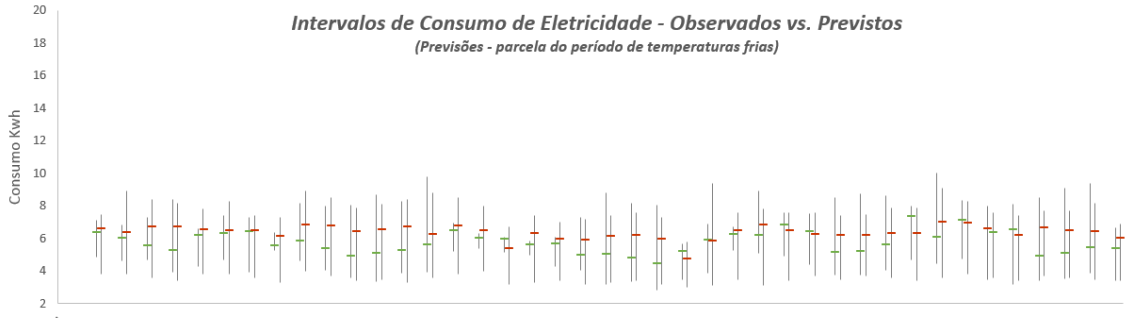


Figura 5.7: Representação dos intervalos observados e previstos para um período frio

Como realizado no estudo anterior, e de forma a comparar os resultados obtidos, também foram aplicados, a este conjunto de dados, os modelos apresentados no Capítulo 3 e o *ID Model*, quando este assume uma distribuição Uniforme ou Triangular Simétrica nos intervalos. A relação linear resultante de cada um dos modelos e os respetivos valores das medidas de qualidade estão apresentados na Tabela 5.5 e na Tabela 5.6, respetivamente

Tabela 5.5: Relações Lineares para cada um dos modelos: aplicação período Quente e período Frio

Modelos	***Expressões que permitem prever os intervalos***
$ID_{Uniforme}$	$\Psi^{-1}_{\widehat{Cons\_Quente}}(t) = 3.7999 + 0.3141\Psi^{-1}_{Temp\_Quente}(t) - 0.0731\Psi^{-1}_{Temp\_Quente}(1-t)$ $\Psi^{-1}_{\widehat{Cons\_Frio}}(t) = 4.2657 + 0.2778\Psi^{-1}_{Temp\_Frio}(t) - 0.0903\Psi^{-1}_{Temp\_Frio}(1-t)$
$ID_{TriangularSimetrica}$	$\Psi^{-1}_{\widehat{Cons\_Quente}}(t) = 3.7999 + 0.3141\Psi^{-1}_{Temp\_Quente}(t) - 0.0731\Psi^{-1}_{Temp\_Quente}(1-t)$ $\Psi^{-1}_{\widehat{Cons\_Frio}}(t) = 4.2657 + 0.2778\Psi^{-1}_{Temp\_Frio}(t) - 0.0903\Psi^{-1}_{Temp\_Frio}(1-t)$
$ID_{Triangular}$	$\Psi^{-1}_{\widehat{Cons\_Quente}}(t) = 3.9218 + 0.3232\Psi^{-1}_{Temp\_Quente}(t) - 0.0699\Psi^{-1}_{Temp\_Quente}(1-t)$ $\Psi^{-1}_{\widehat{Cons\_Frio}}(t) = 4.5607 + 0.2689\Psi^{-1}_{Temp\_Frio}(t) - 0.0845\Psi^{-1}_{Temp\_Frio}(1-t)$
CM	$c_{\widehat{Cons\_Quente}} = 3.799 + 0.2410c_{Temp\_Quente}$ $c_{\widehat{Cons\_Frio}} = 4.2657 + 0.1875c_{Temp\_Frio}$
MinMax	$u_{\widehat{Cons\_Quente}} = 3.6894 + 0.2864u_{Temp\_Quente}$ $l_{\widehat{Cons\_Quente}} = 3.6360 + 0.1583l_{Temp\_Quente}$ $u_{\widehat{Cons\_Frio}} = 5.7323 + 0.1742u_{Temp\_Frio}$ $l_{\widehat{Cons\_Frio}} = 3.2488 + 0.1498l_{Temp\_Frio}$
CRM	$c_{\widehat{Cons\_Quente}} = 3.7999 + 0.2410c_{Temp\_Quente}$ $r_{\widehat{Cons\_Quente}} = 1.7145 + 0.1979r_{Temp\_Quente}$ $c_{\widehat{Cons\_Frio}} = 4.2657 + 0.1875c_{Temp\_Frio}$ $r_{\widehat{Cons\_Frio}} = 1.7992 + 0.0861r_{Temp\_Frio}$
CCRM	$c_{\widehat{Cons\_Quente}} = 3.7999 + 0.2410c_{Temp\_Quente}$ $r_{\widehat{Cons\_Quente}} = 1.7145 + 0.1979r_{Temp\_Quente}$ $c_{\widehat{Cons\_Frio}} = 4.2657 + 0.1875c_{Temp\_Frio}$ $r_{\widehat{Cons\_Frio}} = 1.7992 + 0.0861r_{Temp\_Frio}$

Através dos parâmetros calculados anteriormente, foram obtidos os valores das medidas de qualidade para cada um dos modelos, que estão apresentadas na Tabela 5.6.

Tabela 5.6: Medidas de Qualidade para cada um dos modelos: aplicação período Quente e período Frio

	Modelo Período Quente				***Modelo Período Frio***			
Modelos	$\boldsymbol{\Omega}$	$\boldsymbol{RMSE_M}$	$\boldsymbol{RMSE_L}$	$\boldsymbol{RMSE_U}$	$\boldsymbol{\Omega}$	$\boldsymbol{RMSE_M}$	$\boldsymbol{RMSE_L}$	$\boldsymbol{RMSE_U}$
$ID_{Uniforme}$	0.9419	0.5676	0.5819	0.8823	0.8688	0.5806	0.7548	0.9342
$ID_{TriangularSimetrica}$	0.9399	0.5331	0.6789	0.9431	0.8565	0.4910	0.7548	0.9342
$ID_{Triangular}$	-	0.6754	0.7269	0.9624	-	0.5002	0.9051	0.8778
CM	0.5711	0.7478	1.4422	1.5663	0.8319	0.6279	1.3283	1.4040
$MinMax$	0.8941	0.5159	0.4181	0.7822	0.8768	0.4999	0.4499	0.6516
CRM	0.4790	0.9068	0.5144	0.6252	0.7920	0.6323	0.5527	0.5584
$CCRM$	0.4790	0.9068	0.5144	0.6252	0.7920	0.6323	0.5527	0.5584

Neste estudo observa-se novamente uma igualdade nas estimativas dos parâmetros do modelo $ID_{Uniforme}$ com o $ID_{TriagularSimétrica}$ e do modelo do Centro e Raio com o modelo do Centro e Raio com Restrições.

Analisando a primeira medida de qualidade dos modelos, Ω, os métodos com um pior desempenho, em ambos os modelos, foram o método do Centro e Raio e o método do Centro e Raio com Restrições. É importante realçar que o *ID Model*, para ambas as distribuições, e o modelo do Mínimo e Máximo obtiveram um melhor desempenho no modelo referente ao período quente, onde existe uma maior amplitude dos intervalos, quando comparado com os mesmos valores do modelo relativo ao período frio.

Pela medida de qualidade $RMSE_M$, todos os métodos estudados, com exceção do $ID_{Uniforme,}$ obtiveram um melhor desempenho no modelo do período frio face ao modelo do período quente. Analisando as medidas $RMSE_L$ e $RMSE_U$, o modelo com pior desempenho foi o modelo do Centro, tal como já se tinha verificado no primeiro estudo realizado.

Conclusões dos estudos

Estes dois estudos tiveram como principal objetivo ilustrar o novo modelo desenvolvido no âmbito desta dissertação e a aplicação do código desenvolvido no *R software*. Em ambos os estudos o *ID Model* foi aplicado a intervalos assumindo uma

distribuição Triangular Geral, sendo que o modelo não obteve um desempenho superior ao obtido quando se assume uma distribuição Uniforme ou Triangular Simétrica nos intervalos (Dias e Brito, 2017). No entanto, a possibilidade de se considerar a moda dos intervalos enriquece o modelo e permite que a estrutura dos intervalos, nomeadamente a densidade dos seus valores, seja considerada na regressão linear.

Capítulo 6

Conclusões

A Análise Simbólica de Dados, nomeadamente o estudo das variáveis intervalares, tem sido alvo de grandes desenvolvimentos, impulsionados pelas necessidades de considerar a variabilidade nos dados quando estes são agregados a níveis superiores. Esta necessidade manifesta-se devido à possibilidade em considerar cada vez mais informação na agregação dos dados, demonstrando-se ser algo essencial considerando o panorama atual, em que existe um grande volume disponível.

Têm sido propostos vários métodos de regressão linear para este tipo de variáveis, nesta dissertação foram apresentados os mais relevantes. Este trabalho focou-se no modelo desenvolvido por Dias (2014), onde os intervalos são representados através de funções quantil, permitindo considerar uma maior variabilidade nos dados ao contrário dos restantes modelos desenvolvidos por outros autores. O segundo foco deste trabalho foi a implementação deste modelo para as três distribuições consideradas em *R software*, o que permitiu testar o modelo em bases de dados de larga dimensão e efetuar uma correta avaliação do mesmo, paralelamente com toda a implementação de funções auxiliares, como a distância de *Mallows* e as funções de otimização númerica.

Com o desenvolvimento deste modelo foi também desenvolvida uma função em *R software* que permite realizar a transformação dos dados clássicos em dados simbólicos (variáveis intervalares), agregando-os ao nível desejado e retornando o valor do centro, raio e moda para cada intervalo.

Conclui-se que o desenvolvimento realizado no *ID Model* apresenta bons resultados na previsão dos intervalos, quando comparado com alguns dos métodos já existentes de regressão linear. Este desenvolvimento no modelo tem a vantagem de considerar a moda dos intervalos no modelo de regressão linear. Ao ser incluída a moda é possível ter a visibilidade da densidade dos valores no intervalo, o que enriquece o

desempenho do modelo e a interpretação dos resultados.

No estudo realizado, onde o modelo desenvolvido foi aplicado, os resultados mantiveram um desempenho análogo ao modelo quando é assumida a distribuição Uniforme ou a distribuição Triangular Simétrica.

Apesar do trabalho realizado no âmbito desta dissertação no *ID Model*, existe margem para novos desenvolvimentos no modelo. Um dos desenvolvimentos a estudar é definição da medida de ajuste, Ω, para este modelo quando é assumido uma distribuição Triangular Geral nos intervalos. Com esta medida de avaliação do modelo será possível, com uma maior precisão, avaliar o seu ajuste e comparar o seu desempenho com os restantes modelos. Outra opção de estudo é o desenvolvimento deste modelo para uma relação linear com mais do que uma variável explicativa. Atualmente, a relação estabelecida é apenas entre a variável intervalar resposta e uma variável intervalar explicativa, sendo que acrescentando mais do que uma variável explicativa as estimativas poderão melhorar, nomeadamente a estimativa da moda.

Consideramos que os principais objetivos propostos nesta dissertação foram cumpridos, no entanto, como já mencionado, algumas questões ainda necessitam de ser investigadas.

Bibliografia

Arroyo, Javier, e Carlos Maté. 2009. “Forecasting Histogram Time Series with K-Nearest Neighbours Methods.” *International Journal of Forecasting* 25(1): 192–207. http://www.sciencedirect.com/science/article/pii/S0169207008000678.

Bertoluzza, C., N. C. Blanco e A. Salas. 1995. “On a New Class of Distances between Fuzzy Numbers.” *mathware & Soft Computing* 2(August 2014): 71–84.

Billard, L, e E Diday. 2000. “Regression Analysis for Interval-Valued Data.” In *Data Analysis, Classification, and Related Methods*, eds. Henk A L Kiers, Jean-Paul Rasson, Patrick J F Groenen, and Martin Schader. Berlin, Heidelberg: Springer Berlin Heidelberg, 369–74. https://doi.org/10.1007/978-3-642-59789-3_58.

Billard, Lynne, e Edwin Diday. 2002. “Symbolic Regression Analysis.” In *Classification, Clustering, and Data Analysis: Recent Advances and Applications*, eds. Krzysztof Jajuga, Andrzej Sokołowski, and Hans-Hermann Bock. Berlin, Heidelberg: Springer Berlin Heidelberg, 281–88. https://doi.org/10.1007/978-3-642-56181-8_31.

Bock, Hans Hermann, e E Diday. 2000. Studies in classification, data analysis, and knowledge organization *Analysis of Symbolic Data : Exploratory Methods for Extracting Statistical Information from Complex Data*. http://www.loc.gov/catdir/enhancements/fy0816/99089233-d.html%5Cnhttp://www.loc.gov/catdir/enhancements/fy0816/99089233-t.html.

Brito, Paula. 2014. “Symbolic Data Analysis: Another Look at the Interaction of Data Mining and Statistics.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4(4): 281–95.

Brito, Paula, e M Noirhomme-Fraiture. 2006. “Symbolic and Spacial Data Analysis: Mining Complex Data Structures.” *Guest Editorial, Special Issue of Intelligent Data Analysis* 10: 297–300. http://www.liaad.up.pt/pub/2006/BN06.

Cortez, Paulo, e Aníbal Morais. 2007. “A Data Mining Approach to Predict Forest Fires Using Meteorological Data.” *Proceedings of 13th Portugese Conference on Artificial Intelligence*: 512–23. http://www.dsi.uminho.pt/~pcortez/fires.pdf.

Cheira, Paula; Brito, Paula e Duarte Silva, A. Pedro. 2017. “Factor Analysis of Interval Data.” arXiv:1709.04851 [stat.ME]. Web address: http://arxiv.org/abs/1709.04851

Dias, 2014. Tese de Doutoramento, Universidade do Porto

Dias, Sónia, e Paula Brito. 2017. “Off the Beaten Track: A New Linear Model for Interval Data.” *European Journal of Operational Research* 258(3): 1118–30.

Giordani, Paolo. 2014. “Lasso-Constrained Regression Analysis for Interval-Valued Data.” *Advances in Data Analysis and Classification* 9(1): 5–19.

Hofer, Vera. 2015. “Adapting a Classification Rule to Local and Global Shift When Only Unlabelled Data Are Available.” *European Journal of Operational Research* 243(1): 177–89. http://www.sciencedirect.com/science/article/pii/S037722171400945X.

Irpino, Antonio, e Rosanna Verde. 2006. “A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data.” In *Data Science and Classification*, eds. Vladimir Batagelj, Hans-Hermann Bock, Anuška Ferligoj, and Aleš Žiberna. Berlin, Heidelberg: Springer Berlin Heidelberg, 185–92. https://doi.org/10.1007/3-540-34416-0_20.

Irpino, Antonio, e Rosanna Verde. 2015. “Basic Statistics for Distributional Symbolic Variables: A New Metric-Based Approach.” *Advances in Data Analysis and Classification* 9(2): 143–75. https://doi.org/10.1007/s11634-014-0176-4.

Lima Neto, E. de A, eF. de A T de Carvalho. 2008. “Centre and Range Method for Fitting a Linear Regression Model to Symbolic Interval Data.” *Computational Statistics and Data Analysis* 52(3): 1500–1515.

Lima Neto, Eufrásio de A, e Francisco de A T de Carvalho. 2010. “Constrained Linear Regression Models for Symbolic Interval-Valued Variables.” *Computational Statistics and Data Analysis* 54(2): 333–47.

http://dx.doi.org/10.1016/j.csda.2009.08.010.

Mallows, C. L. 1972. “A Note on Asymptotic Joint Normality.” *The Annals of Mathematical Statistics* 43(2): 508–15. http://projecteuclid.org/euclid.aoms/1177692631%5Cnhttps://projecteuclid.org/download/pdf_1/euclid.aoms/1177692631%5Cnhttps://projecteuclid.org/euclid.aoms/1177692631.

ANEXOS

Função TOSIMBOLIC

```
ToSimbolic <- function(X,m) {
  agregador <- unique(X[,1])
  if(m == 1 | m == 2) {
    simbolic <- matrix(data = 0, nrow = length(agregador), ncol = 2)
    for(i in 1:length(agregador)) {
      data <- 0
      data <- X[X[,1] == agregador[i],]
      simbolic[i,1] <- (max(data[,2]) + min(data[,2]))/2
      simbolic[i,2] <- (max(data[,2]) - min(data[,2]))/2
    }
    return(simbolic)
  } else if(m == 3) {
    simbolic <- matrix(data = 0, nrow = length(agregador), ncol =  3)
    for(i in 1:length(agregador)) {
      data <- 0
      data <- X[X[,1] == agregador[i],]
      simbolic[i,1] <- (max(data[,2]) + min(data[,2]))/2
      simbolic[i,2] <- (max(data[,2]) - min(data[,2]))/2
      if(nrow(data)==1) {simbolic[i,3] <- (max(data[,2]) + min(data[,2]))/2}
      else{simbolic[i,3] <- getMode(data[,2])}
    }
    return(simbolic)
  } else {
    return("Escolher 1 para utilizar o modelo assumindo uma distribuição uniforme, 2 para utilizar o
modelo assumindo uma distribuição triangular simétrica ou 3 para utilizar o modelo assumindo uma
distribuição triangular geral")
  }
}

ToSimbolic2 <- function(X,m) {
  var <- ncol(X)-1
  A <- NULL
  for(i in 1:var) {
    D <- ToSimbolic(X[,c(1,i+1)],m)
    A <- cbind(A,D)
  }
  return(A)
}

getMode <- function(x){
  xdens <- density(x)
  modex <- xdens$x[which.max(xdens$y)]
  return(modex)
}
```

Função Compilador:

```
Modelo <- function(Model, Y, X, Xe,p = c(0.5,0.5,0.5)) {
  if(Model == 1) {
    RL <- RLParametrosModelo1(Y, X)
    Ye <- RLEstimar(Xe, RL)
    Metricas <- RLQualidade(Model, Y, Ye)
    ModeloFinal <-list("Parametros do Modelo" = RL,"Intervalos Estimados" = Ye,  "Medidas de
Qualidade do Modelo" = Metricas)
    return(ModeloFinal)
```

```
  } else if(Model == 2) {
    RL <- RLParametrosModelo2(Y, X)
    Ye <- RLEstimar(Xe, RL)
    Metricas <- RLQualidade(Model, Y, Ye)
    ModeloFinal <-list("Parametros do Modelo" = RL,"Intervalos Estimados" = Ye,  "Medidas de
Qualidade do Modelo" = Metricas)
    return(ModeloFinal)
  } else if(Model == 3) {
    RL <- RLParametrosModelo3(Y,X)$par
    Ye <- Modelo3Previsao(Xe,RL)
    Metricas <- RLQualidade2(Y,Ye)
    ModeloFinal <-list("Parametros do Modelo" = RL," Intervalos Estimados" = Ye,  "Medidas de
Qualidade do Modelo" = Metricas)
    return(ModeloFinal)
  }
}
```

Função Distância de Mallows

```
DistMallows <- function(Model, A, B) {
  Centros <- cbind(as.matrix(A[,1]), as.matrix(B[,1]))
  Raios   <- cbind(as.matrix(A[,2]), as.matrix(B[,2]))
  C <- matrix(rep(0,x = nrow(Centros)))
  R <- matrix(rep(0, nrow(Raios)))
  for(i in 1:nrow(Centros)) {
    C[i, ] <- (Centros[i, 1] - Centros[i, 2])^2
  }
  for(i in 1:nrow(Raios)) {
    if(Model == 1) {
      R[i, ] <- (1 / 3) * (Raios[i, 1] - Raios[i, 2])^2
    } else if (Model == 2) {
      R[i, ] <- (1 / 6) * (Raios[i, 1] - Raios[i, 2])^2
    } else {
      print("escolher modelo 1-Uniforme, 2-Triangular Simetrico, 3-Triangular Geral")
    }
  }
  M  <- C + R
  DM <- sum(M[, 1])
  return(DM)
}
```

Função Parâmetros ID Model Uniforme:

```
RLParametrosModelo1 <- function(Y, X) {
  Y <- as.matrix(Y)
  X <- as.matrix(X)
  Yc <- as.matrix(Y[, 1])
  Yr <- as.matrix(Y[, 2])
  Xc <- as.matrix(X[, seq(from = 1, to = ncol(X), by = 2)])
  Xr <- as.matrix(X[, seq(from = 2, to = ncol(X), by = 2)])
  p  <- ncol(Xc)

  NHhessianaBlocoI <- function(Y,X) {
    H   <- matrix(0, p * 2, p * 2)
    Xcc <- crossprod(Xc)
    Xrr <- crossprod(Xr)
```

```
  for(k in 1:(p * 2)) {
    for(l in 1:(p * 2)) {
      if(k %% 2 != 0) {
        H <- H
      } else if(l %% 2 != 0) {
        H <- H
      } else {
        H[k, l]     <-  2 * Xcc[k/2, l/2] + 2/3 * Xrr[k/2, l/2]
        H[k-1, l-1] <-  2 * Xcc[k/2, l/2] + 2/3 * Xrr[k/2, l/2]
        H[k-1, l]   <-  -2 * Xcc[k/2, l/2] + 2/3 * Xrr[k/2, l/2]
        H[k, l-1]   <-  -2 * Xcc[k/2, l/2] + 2/3 * Xrr[k/2, l/2]
      }
    }
  }
  H <- round(H, 2)
  return(H)
}

H1 <- NHhessianaBlocoI(Y, X)

NHhessianaBlocoLinha <- function(Y, X) {
  HA <- matrix(0, 1, p * 2)
  for(l in 1:(p * 2)) {
    if(l %% 2 != 0) {
      HA <- HA
    } else {
      HA[, l]   <- -2 * sum(Xc[, l/2])
      HA[, l-1] <- 2 * sum(Xc[, l/2])
    }
  }
  HA <- list("down" = HA, "side" = t(HA), "constante" = nrow(Xc) * 2)
  return(HA)
}

HA <- NHhessianaBlocoLinha(Y, X)

NHhessiana <- function(Y, X) {
  NH <- cbind(rbind(H1, HA$down), rbind(HA$side, HA$constante))
  return(NH)
}
H <- NHhessiana(Y, X)

NHfuncao <- function(Y, X) {
  Hf  <- matrix(0, p * 2, 1)
  Cf  <- matrix(0, 1, 1)
  YXc <- matrix(0, p, 1)
  YXr <- matrix(0, p, 1)
  for(i in 1:p) {
    YXc[i, 1] <- crossprod(Yc[, 1], Xc[, i])
    YXr[i, 1] <- crossprod(Yr[, 1], Xr[, i])
  }
  for(l in 1:p * 2) {
    if(l %% 2 != 0) {
      Hf <- Hf
    } else {
      Hf[l, ]   <- 2 * YXc[l/2, ] - 2/3 * YXr[l/2, ]
      Hf[l-1, ] <- -2 * YXc[l/2, ] - 2/3 * YXr[l/2, ]
    }
```

```
    }
    Cf[1, 1] <- -2 * sum(Yc)
    w1 <- rbind(Hf, Cf)
    return(w1)
  }
  F <- NHfuncao(Y, X) * (-1)

library("quadprog")
  A  <- matrix(0, p*2, p*2)
  diag(A) <- 1
  A  <- rbind(A, matrix(0, 1, p*2))
  b  <- matrix(0, p*2, 1)
  RL <- solve.QP(H, F, A, b)
  return(RL$solution)
}
```

Função Parâmetros ID Model Triangular Simétrica:

```
RLParametrosModelo2 <- function(Y, X) {
  Y <- as.matrix(Y)
  X <- as.matrix(X)
  Yc <- as.matrix(Y[, 1])
  Yr <- as.matrix(Y[, 2])
  Xc <- as.matrix(X[, seq(from = 1, to = ncol(X), by = 2
  Xr <- as.matrix(X[, seq(from = 2, to = ncol(X), by = 2)])
  p  <- ncol(Xc)

  NHhessianaBlocoI <- function(Y, X) {
    H   <- matrix(0, p * 2, p * 2)
    Xcc <- crossprod(Xc)
    Xrr <- crossprod(Xr)
    for(k in 1:(p * 2)) {
      for(l in 1:(p * 2)) {
        if(k %% 2 != 0) {
          H <- H
        } else if(l %% 2 != 0) {
          H <- H
        } else {
          H[k, l]     <-  2 * Xcc[k/2, l/2] + 1/3 * Xrr[k/2, l/2]
          H[k-1, l-1] <-  2 * Xcc[k/2, l/2] + 1/3 * Xrr[k/2, l/2]
          H[k-1, l]   <-  -2 * Xcc[k/2, l/2] + 1/3 * Xrr[k/2, l/2]
          H[k, l-1]   <-  -2 * Xcc[k/2, l/2] + 1/3 * Xrr[k/2, l/2]
        }
      }
    }
    H <- round(H, 2)
    return(H)
  }
  H1 <- NHhessianaBlocoI(Y, X)

NHhessianaBlocoLinha <- function(Y, X) {

    HA <- matrix(0, 1, p * 2)
    for(l in 1:(p * 2)) {
      if(l %% 2 != 0) {
        HA <- HA
      } else {
```

```
      HA[, l]   <- -2 * sum(Xc[, l/2])
      HA[, l-1] <- 2 * sum(Xc[, l/2])
    }
  }
  HA <- list("down" = HA , "side" = t(HA) , "constante" = nrow(Xc) * 2)
  return(HA)
 }
 HA <- NHhessianaBlocoLinha(Y, X)

 NHhessiana <- function(Y, X) {
  NH <- cbind(rbind(H1, HA$down), rbind(HA$side, HA$constante))
  return(NH)
 }
 H <- NHhessiana(Y, X)

 NHfuncao <- function(Y, X) {
  Hf  <- matrix(0, p * 2, 1)
  Cf  <- matrix(0, 1, 1)
  YXc <- matrix(0, p, 1)
  YXr <- matrix(0, p, 1)
  for(i in 1:p) {
    YXc[i, 1] <- crossprod(Yc[, 1], Xc[, i])
    YXr[i, 1] <- crossprod(Yr[, 1], Xr[, i])
  }
  for(l in 1:p * 2) {
    if(l %% 2 != 0) {
      Hf <- Hf
    } else {
      Hf[l, ]   <- 2 * YXc[l/2, ] - 1/3 * YXr[l/2, ]
      Hf[l-1, ] <- -2 * YXc[l/2, ] - 1/3 * YXr[l/2, ]
    }
  }
  Cf[1, 1] <- -2 * sum(Yc)
  w1       <- rbind(Hf, Cf)
  return(w1)
 }
 F <- NHfuncao(Y, X) * (-1)

 library("quadprog")
 A <- matrix(0, p * 2, p * 2)
 diag(A) <- 1
 A <- rbind(A, matrix(0, 1, p * 2))
 b <- matrix(0, p * 2, 1)
 RL <- solve.QP(H, F, A, b)
 return(RL$solution)
}
```

Função Parâmetros ID Model Triangular Geral:

```
RLParametrosModelo3 <- function(Y, X,p) {
  initialRL <- p
  RLvar <- c(99,99,99)
  SolOpt <- RepLOptim(parmean = initialRL,parsd = RLvar, fr = OptFuction,lower = c(0,0,-Inf), Y = Y,
X = X)
  return(SolOpt)
}
```

Função Minimizadora na Otimização Númerica

```
OptFuction <- function(initialRL,Y,X) {
  Y <- as.matrix(Y)
  X <- as.matrix(X)
  YC <- as.matrix(Y[, 1])
  YR <- as.matrix(Y[, 2])
  YM <- as.matrix(Y[, 3])
  XC <- as.matrix(X[, 1])
  XR <- as.matrix(X[, 2])
  XM <- as.matrix(X[, 3])
  p  <- ncol(XC)
  n  <- nrow(X)
  alfa  <- initialRL[1]
  beta  <- initialRL[2]
  gama  <- initialRL[3]
  try(if(p > 1) stop("model only supports one variable"))
    sum <- 0
  for(i in 1:n) {
    Yc <- 0
    Yr <- 0
    Ym <- 0
    Xc <- 0
    Xr <- 0
    Xm <- 0
    Yc <- YC[i, ]
    Yr <- YR[i, ]
    Ym <- YM[i, ]
    Xc <- XC[i, ]
    Xr <- XR[i, ]
    Xm <- XM[i, ]
    Xa <- Xc - Xr
    Xb <- Xc + Xr
    Ymest <- 0
    Yaest <- 0
    Ybest <- 0
    Yrest <- 0
    Ycest <- 0
    if(round((Xm - Xa),4) > round((Xb - Xm),4)) {
      if((alfa > ((beta * (Xb - Xm)^(3 / 2)) / (Xm - Xa)^(3 / 2))) & (alfa < ((beta * (Xm - Xa)^(3 / 2)) / (Xb -
Xm)^(3 / 2)))) {
        Ymest <- (alfa-beta)*Xa + sqrt((Xb-Xa)*(Xm-Xa))*((alfa^(4/3)-
beta^(4/3))/sqrt(alfa^(2/3)+beta^(2/3))) + gama
        Yaest <- alfa * Xa - beta * Xb + gama
        Ybest <- alfa * Xb - beta * Xa + gama
        Ycest <- (Ybest + Yaest)/2
        Yrest <- (Ybest - Yaest)/2
      } else if(alfa <= ((beta * (Xb - Xm) ^ (3 / 2)) / (Xm - Xa)^(3 / 2))) {
        Ymest <- alfa * Xa - beta * Xm + alfa * sqrt((Xb-Xm) * (Xm-Xa)) + gama
        Yaest <- alfa * Xa - beta * Xb + gama
        Ybest <- alfa * Xb - beta * Xa + gama
        Ycest <- (Ybest + Yaest)/2
        Yrest <- (Ybest - Yaest)/2
      } else if(alfa >= ((beta * (Xm - Xa)^(3 / 2)) / (Xb - Xm)^(3 / 2))) {
        Ymest <- alfa * Xm - beta * Xa + beta * sqrt((Xb-Xm) * (Xm-Xa)) + gama
        Yaest <- alfa * Xa - beta * Xb + gama
        Ybest <- alfa * Xb - beta * Xa + gama
        Ycest <- (Ybest + Yaest)/2
```

```
      Yrest <- (Ybest - Yaest)/2
     }
   } else if(round((Xm - Xa),4) < round((Xb - Xm),4)) {
    if((alfa > ((beta * (Xm - Xa)^(3 / 2)) / (Xb - Xm)^(3 / 2))) & (alfa < ((beta * (Xb - Xm)^(3 / 2)) / (Xm
- Xa)^(3 / 2)))) {
      Ymest <- (alfa-beta)*Xb + sqrt((Xb-Xa)*(Xb-Xm))*((beta^(4/3)-
alfa^(4/3))/sqrt(alfa^(2/3)+beta^(2/3))) + gama
      Yaest <- alfa * Xa - beta * Xb + gama
      Ybest <- alfa * Xb - beta * Xa + gama
      Ycest <- (Ybest + Yaest)/2
      Yrest <- (Ybest - Yaest)/2
     } else if(alfa <= ((beta * (Xm - Xa)^(3 / 2)) / (Xb - Xm)^(3 / 2))) {
      Ymest <- alfa * Xb - beta * Xm + alfa * sqrt((Xb-Xm) * (Xm-Xa)) + gama
      Yaest <- alfa * Xa - beta * Xb + gama
      Ybest <- alfa * Xb - beta * Xa + gama
      Ycest <- (Ybest + Yaest)/2
      Yrest <- (Ybest - Yaest)/2
     } else if(alfa >= ((beta * (Xb - Xm)^(3 / 2)) / (Xm - Xa)^(3 / 2))) {
      Ymest <- alfa * Xm - beta * Xb + alfa * sqrt((Xb-Xm) * (Xm-Xa)) + gama
      Yaest <- alfa * Xa - beta * Xb + gama
      Ybest <- alfa * Xb - beta * Xa + gama
      Ycest <- (Ybest + Yaest)/2
      Yrest <- (Ybest - Yaest)/2
     }
   } else if(round((Xm - Xa),4) == round((Xb - Xm),4)) {
     Yaest <- alfa * Xa - beta * Xb + gama
     Ybest <- alfa * Xb - beta * Xa + gama
     Ycest <- (Ybest + Yaest)/2
     Yrest <- (Ybest - Yaest)/2
     Ymest <- Ycest
   }
  DM <- DistMallows2(Yc,Yr,Ym,Ycest,Yrest,Ymest)
  sum <- sum + DM
  }
  return(sum)
}
```

Função Estimar

```
RLEstimar <- function(Xe, RL) {
  Xe <- as.matrix(Xe)
  parametros <- matrix(RL[-length(RL)], (length(RL) - 1) / 2, 2, byrow = TRUE)
  parametros_raios <- as.matrix(parametros[, 1] + parametros[, 2])
  parametros_centros <- rbind(as.matrix(parametros[, 1] - parametros[, 2]), RL[length(RL)])
  Xc <- as.matrix(Xe[, seq(from = 1, to = ncol(Xe), by = 2)])
  Xr <- as.matrix(Xe[, seq(from = 2, to = ncol(Xe), by = 2)])
  Yec <- cbind(Xc, 1) %*% parametros_centros
  Yer <- cbind(Xr)    %*% parametros_raios
  return(cbind(Yec, Yer))
}
Modelo3Previsao <- function(Xe,RL) {
  alfa <- RL[1]
  beta <- RL[2]
  gama <- RL[3]
  X  <- as.matrix(Xe)
  n  <- nrow(X)
  Ye <- matrix(0, n, 3)
```

```
  XC <- as.matrix(X[, 1])
  XR <- as.matrix(X[, 2])
  XM <- as.matrix(X[, 3])
  for(i in 1:n) {
    Xc <- 0
    Xr <- 0
    Xm <- 0
    Xc <- XC[i, ]
    Xr <- XR[i, ]
    Xm <- XM[i, ]
    Xa <- Xc - Xr
    Xb <- Xc + Xr
    Ymest <- 0
    Yaest <- 0
    Ybest <- 0
    Yrest <- 0
    Ycest <- 0
    if(round((Xm - Xa),4) > round((Xb - Xm),4)) {
      if((alfa > ((beta * (Xb - Xm)^(3 / 2)) / (Xm - Xa)^(3 / 2))) & (alfa < ((beta * (Xm - Xa)^(3 / 2)) / (Xb -
Xm)^(3 / 2)))) {
        Ye[i,3] <- (alfa-beta)*Xa + sqrt((Xb-Xa)*(Xm-Xa))*((alfa^(4/3)-
beta^(4/3))/sqrt(alfa^(2/3)+beta^(2/3))) + gama
        Yaest   <- alfa * Xa - beta * Xb + gama
        Ybest   <- alfa * Xb - beta * Xa + gama
        Ye[i,1] <- (Ybest + Yaest)/2
        Ye[i,2] <- (Ybest - Yaest)/2
      } else if(alfa < ((beta * (Xb - Xm) ^ (3 / 2)) / (Xm - Xa)^(3 / 2))) {
        Ye[i,3] <- alfa * Xa - beta * Xm + alfa * sqrt((Xb-Xm) * (Xm-Xa)) + gama
        Yaest   <- alfa * Xa - beta * Xb + gama
        Ybest   <- alfa * Xb - beta * Xa + gama
        Ye[i,1] <- (Ybest + Yaest)/2
        Ye[i,2] <- (Ybest - Yaest)/2
      } else if(alfa > ((beta * (Xm - Xa)^(3 / 2)) / (Xb - Xm)^(3 / 2))) {
        Ye[i,3] <- alfa * Xm - beta * Xa + beta * sqrt((Xb-Xm) * (Xm-Xa)) + gama
        Yaest   <- alfa * Xa - beta * Xb + gama
        Ybest   <- alfa * Xb - beta * Xa + gama
        Ye[i,1] <- (Ybest + Yaest)/2
        Ye[i,2] <- (Ybest - Yaest)/2
      }
    } else if(round((Xm - Xa),4) < round((Xb - Xm),4)) {
      if((alfa > ((beta * (Xm - Xa)^(3 / 2)) / (Xb - Xm)^(3 / 2))) & (alfa < ((beta * (Xb - Xm)^(3 / 2)) / (Xm
- Xa)^(3 / 2)))) {
        Ye[i,3] <- (alfa-beta)*Xb + sqrt((Xb-Xa)*(Xb-Xm))*((beta^(4/3)-
alfa^(4/3))/sqrt(alfa^(2/3)+beta^(2/3))) + gama
        Yaest   <- alfa * Xa - beta * Xb + gama
        Ybest   <- alfa * Xb - beta * Xa + gama
        Ye[i,1] <- (Ybest + Yaest)/2
        Ye[i,2] <- (Ybest - Yaest)/2
      } else if(alfa < ((beta * (Xm - Xa)^(3 / 2)) / (Xb - Xm)^(3 / 2))) {
        Ye[i,3] <- alfa * Xb - beta * Xm + alfa * sqrt((Xb-Xm) * (Xm-Xa)) + gama
        Yaest   <- alfa * Xa - beta * Xb + gama
        Ybest   <- alfa * Xb - beta * Xa + gama
        Ye[i,1] <- (Ybest + Yaest)/2
        Ye[i,2] <- (Ybest - Yaest)/2
      } else if(alfa > ((beta * (Xb - Xm)^(3 / 2)) / (Xm - Xa)^(3 / 2))) {
        Ye[i,3] <- alfa * Xm - beta * Xb + alfa * sqrt((Xb-Xm) * (Xm-Xa)) + gama
        Yaest   <- alfa * Xa - beta * Xb + gama
        Ybest   <- alfa * Xb - beta * Xa + gama
```

```
      Ye[i,1] <- (Ybest + Yaest)/2
      Ye[i,2] <- (Ybest - Yaest)/2
    }
   } else if(round((Xm - Xa),4) == round((Xb - Xm),4)) {
    Yaest   <- alfa * Xa - beta * Xb + gama
    Ybest   <- alfa * Xb - beta * Xa + gama
    Ye[i,1] <- (Ybest + Yaest)/2
    Ye[i,2] <- (Ybest - Yaest)/2
    Ye[i,3] <- (Ybest + Yaest)/2
   }
  }
  return(Ye)
}
```

Função Qualidade

```
RLQualidade <- function(Model, Y, Ye) {
  Y  <- as.matrix(Y)
  Ye <- as.matrix(Ye)
  Ym    <- matrix(rep(mean(Y[, 1]), nrow(Y)))
  Ymean <- cbind(Ym, 0)
  DMe     <- DistMallows(Model, Ye, Ymean)
  DM      <- DistMallows(Model, Y, Ymean)
  DMrmsem <- DistMallows(Model, Y, Ye)
  CD <- DMe / DM
  Yeu <- Ye[, 1] + Ye[, 2]
  Yel <- Ye[, 1] - Ye[, 2]
  Yru <- Y[, 1] + Y[, 2]
  Yrl <- Y[, 1] - Y[, 2]
  A1 <- Yru - Yeu
  A2 <- Yrl - Yel
  B1 <- (mean(A1^2))
  B2 <- (mean(A2^2))
  RMSEM <- sqrt(DMrmsem / nrow(Ye))
  RMSEU <- B1^0.5
  RMSEL <- B2^0.5
  return(list(c("CD"= CD, "RMSEM" = RMSEM, "RMSEL" = RMSEL, "RMSEU" = RMSEU)))
}
RLQualidade2 <- function(Y, Ye) {
  Y  <- as.matrix(Y)
  Cy <- Y[,1]
  Ry <- Y[,2]
  My <- Y[,3]
  Ye <- as.matrix(Ye)
  Cye <- Ye[,1]
  Rye <- Ye[,2]
  Mye <- Ye[,3]
  n  <- nrow(Ye)
  DistMallFinal <- 0
  for (i in 1:n) {
    DM <- DistMallows2(Cye[i],Rye[i],Mye[i],Cy[i],Ry[i],My[i])
    DistMallFinal <- DistMallFinal + DM
  }
  Yeu <- Ye[, 1] + Ye[, 2]
  Yel <- Ye[, 1] - Ye[, 2]
  Yru <- Y[, 1] + Y[, 2]
  Yrl <- Y[, 1] - Y[, 2]
```

```
 A1 <- Yru - Yeu
 A2 <- Yrl - Yel
 B1 <- (mean(A1^2))
 B2 <- (mean(A2^2))
 RMSEM <- sqrt(DistMallFinal/ nrow(Ye))
 RMSEU <- B1^0.5
 RMSEL <- B2^0.5
 return(list(c("RMSEM" = RMSEM, "RMSEL" = RMSEL, "RMSEU" = RMSEU)))
}
```